Annals of Operations Research
https://doi.org/10.1007/s10479-022-04880-4

ORIGINAL RESEARCH

Forecast combinations for benchmarks of long-term stock returns using machine learning methods

Michael Scholz[1]

Accepted: 15 July 2022

Abstract
Forecast combinations are a popular way of reducing the mean squared forecast error when multiple candidate models for a target variable are available. We apply different approaches to finding (optimal) weights for forecasts of stock returns in excess of different benchmarks. Our focus lies thereby on nonlinear predictive functions estimated by a fully nonparametric smoother with the covariates and the smoothing parameters chosen by cross-validation. Based on an out-of-sample study, we find that individual nonparametric models outperform their forecast combinations. The latter are prone to in-sample over-fitting and in consequence, perform poorly out-of-sample especially when the set of possible candidates for combinations is large. A reduction to one-dimensional models balances in-sample and out-of-sample performance.

Keywords Forecasting · Machine learning · Forecast combinations · Nonlinear prediction · Stock returns

1 Introduction

Financial investment planning for long-term savings is highly relevant for the development of new pension products (Merton, 2014; Gerrard et al., 2019, 2020). Therefore, understanding the dynamics of the stock market is crucial in providing the long-term saver with sufficient wealth at retirement. It is well-known from the empirical literature that model-based predictions for longer horizons can provide better forecasts than the simple historical mean (Campbell & Thompson, 2008). However, a careful validation approach has to be applied when predictions of stock returns are based on reasonable long-term economic drivers. In this paper, we focus on nonlinear predictive functions which are estimated with a fully data-driven local-linear smoother in combination with a leave-k-out cross-validation for the prediction of stock returns in excess of different benchmarks as developed in Kyriakou et al. (2021a). These functions optimally incorporate the given information as they allow for complex interrelations of the potential predictor variables. We work with low-dimensional models and estimate individually for each selection of variables the nonlinear predictive relationship. However,

✉ Michael Scholz
michael.scholz@aau.at

[1] Department of Economics, University of Klagenfurt, Klagenfurt, Austria

Published online: 02 August 2022

forecast combinations are known to potentially reduce the mean squared forecast error when several individual candidates are available. Thus, we not only validate the predictive power of the individual forecasts but also analyse whether it is beneficial to combine them in several ways. Recently, machine learning (ML) algorithms have been proposed for this purpose and we focus mainly on weighting schemes for the forecast combinations which are based on such techniques like the Lasso, the Ridge, and the Elastic Net, as well as their egalitarian variants, or recently introduced refinements (Combination Elastic Net). We employ historical S&P500 returns in excess of different benchmarks, including the short-term interest rate and the inflation, at the annual frequency for a sample period ranging from 1872 to 2022.

The contributions of this paper are manifold. First, we extend the nonlinear prediction framework of Kyriakou et al. (2021a) to also considering three-dimensional models. We show that such complex models can have reasonable predictive power both in-sample and out-of-sample. For example, under the short-term interest rate benchmark, three of the five models with the largest out-of-sample predictive power are three-dimensional. Thereby, the model based on time-lagged excess returns, dividends, and term spread is the second-best predictive model for the risk-premium (in terms of a large out-of-sample R^2 value). Under the inflation benchmark, four of the five models with the largest out-of-sample predictive power are three-dimensional. The model based on real-dividends, real-earnings, and term-spread performed best in predicting real stock returns out-of-sample (cf. Tables 2, 6). Second, we find that individual nonparametric forecasts usually outperform forecast combination methods based on ML techniques. If we allow under the short-term interest rate benchmark only for one-dimensional candidates, then the forecast combinations give slightly better predictions than the best individual model. Thus, the complexity introduced in the prediction process when using ML-based techniques does not pay off well enough and it is better to use simpler and more transparent methods. Third, we highlight that the classical shrinkage methods are prone to in-sample over-fitting when too many individual forecasts are used as possible candidates. The consequence is that the suggested predictive power is spurious and the out-of-sample performance very poor. However, using only the one-dimensional models balances in-sample and out-of-sample behaviour. Note also that forecast combinations perform better than the simple historical average. Fourth, considering all variables in real terms net of inflation (the inflation double benchmark) results in a much more stable and consistent analysis both between models and over time when compared to the prediction of the risk premium (short-term interest rate single benchmark). This is especially important for the long-term pension saver who is interested in adequate strategies for real-income protection.

The remaining of this paper is organized as follows. Section 2 presents the literature review. In Sect. 3, we introduce our long-term predictive framework, outline the estimation procedure using the local-linear smoother, describe different ways of combing the individual forecasts, and give an overview of the US stock market data. Section 4 provides a discussion about the results of the empirical study for the prediction of the risk-premium and real returns. Section 5 summarizes the key points of our analysis and concludes the paper.

2 Literature review

In the last decades, numerous studies in the academic literature focus on answering the question of whether asset returns are predictable or not. From an economic perspective, it was commonly assumed until the mid-1980s that predictability would contradict the efficient markets hypothesis (Fama, 1970). However, the seminal work by Fama (1988), Campbell and

Shiller (1988), or Stambaugh (1999) suggests the nowadays 'common wisdom' of long term predictability (Lioui & Poncet, 2019). For more recent approaches regarding stock market forecasts, see, for example, Scholz et al. (2015), Scholz et al. (2016), Lioui and Poncet (2019), or Akyildirim et al. (2022) and the discussion therein.

From the statistical or econometric point of view, the prediction setup can be described in the following very general way (Hastie et al., 2017):

$$\min_{f\in\mathcal{H}}\left\{L\big(y_{t+h},\, f(Z_t)\big)+p(f,\tau)\right\},\quad t=1,\ldots,T, \tag{1}$$

where y_{t+h} is the variable to be predicted h periods ahead, Z_t the vector of predictors, $\mathcal{H}$ a space of possible functions f that combine the data to form the prediction, p a penalty on f, τ a set of hyper-parameters (for example, the λ in the Lasso), and L a loss function that defines the optimal forecast.

In this article, we take the long-term actuarial perspective and base our empirical study on annual observations. Thus, we are not in a big-data context where the number of observations is huge. The set of possible predictive variable combinations is also rather small. In other words, we can work with low dimensional models in (1), and shrinkage, dimension reduction, or penalization are not necessary. However, sparsity could be an issue with our data set and a careful imposition of structure in the statistical modelling process is helpful. Note further that the use of nonlinear functions f in (1) has shown evidence of much stronger stock return predictability when compared to their linear counterparts (Lettau & Van Nieuwerburgh, 2008; Chen & Hong, 2010; Yang et al., 2010; Cheng et al., 2019; Caldeira et al., 2020; Freyberger et al., 2020). Thus, the local-linear smoother based on the standard L_2-loss function is ideally suited. Note that a linear function—the classical benchmark in this context—can be estimated without any bias.

Several studies are based on this technique. Most of them try to improve the prediction utilizing additional structure in the estimation process and to reduce the impact of the curse of dimensionality in a sparse data environment. Nielsen and Sperlich (2003) were the first to introduce this nonparametric technique together with an adequate validation method into the actuarial literature. Scholz et al. (2015) use bootstrap techniques to formally test the null hypothesis of the non-predictability of returns and improve the smoothing through prior knowledge using a multiplicative bias-reduction approach. Scholz et al. (2016) propose a two-step procedure for the prediction of excess stock returns: (i) same-years bond yield is constructed fully nonparametrically, and (ii) this additional predictor is used to forecast excess stock returns. Mammen et al. (2019) focus on the prediction of the conditional variance of long-term stock returns. They find that volatility forecastability is much less important at longer horizons and that the homoscedastic historical average of the squared return prediction errors give adequate approximations of the unobserved realised conditional variance. Kyriakou et al. (2020) consider the 5-year horizon and corresponding econometric challenges like overlapping observations. They find that long-term forecasting performs well and recommend drawing more attention to it when designing investment strategies for long-term investors. Kyriakou et al. (2021a) propose the use of different benchmarks when predicting stock returns. Their full benchmarking approach, that is, considering all variables net of inflation, has important consequences for long-term saving strategies, where one is interested in real value. Finally, Kyriakou et al. (2021b) propose an econometric model which combines different horizons. Their method exploits the lower long-term variance to further reduce the short-term variance, which is susceptible to speculative exuberance. As a consequence, the long-term pension-saver avoids an over-conservative portfolio with implied potential upside reductions given their optimal risk appetite. Our study analyses now the question of whether

the combination of individual forecasts based on ML techniques can improve predictability as it was recently documented in the literature, for example, by Rapach and Zhou (2020). However, we find that this kind of complexity does not pay off well enough and we recommend the use of simpler individual forecasts.

ML is one of the in-vogue topics in empirical finance and actuarial science (Asimit et al., 2020; Dixon et al., 2020) for asset return prediction or portfolio choice (Coqueret & Guida, 2020; Akyildirim et al., 2021, 2022). It is often seen as " (i) a diverse collection of high-dimensional models for statistical prediction, combined with (ii) so-called 'regularization' methods for model selection and mitigation of overfit, and (iii) efficient algorithms for searching among a vast number of potential model specifications" (Gu et al., 2020). Mostly one of the following methods is well suited to address the three challenges mentioned earlier: linear models for regression (including regularization via shrinkage methods with penalization, such as Ridge Regression, Lasso, or Elastic Nets), dimension reduction via principal components regression and partial least squares, regression trees and forests (including boosted trees and random forests), (deep) neural networks, and boosting (Oztekin et al., 2016; Athey & Imbens, 2019; Coulombe et al., 2020; Gu et al., 2020; Hiabu et al., 2020; Iworiso & Vrontos, 2020; Wu et al., 2020; Gambella et al., 2021).

Forecast combinations are a popular way of reducing the mean squared forecast error when several individual predictive models (usually of low dimensionality) for a target variable are available. The forecasting ability of individual predictive regression models could be seriously impaired by model uncertainty and (parameter) instability (Rapach et al., 2010). Several methods of finding the (optimal) combination forecast have been proposed in a large body of literature: for example, a weighted average of forecasts, with the weights adding up to unity (Granger & Ramanathan, 1984); trimming (Granger & Jeon, 2004); rank-based approaches (Aiolfi & Timmermann, 2006); a least-squares forecast averaging (Hansen, 2008b); a complete subset regression (Elliott et al., 2013); iterated (Lin et al., 2018) or depth-weighted combinations (Lee & Sul, 2021). Recently, ML techniques have been proposed to select and weight appropriate individual forecasts using, for example, Lasso-based procedures (Diebold & Shin, 2019; Mascio et al., 2020; Freyberger et al., 2020); a combining method for sophisticated models with the historical average serving as shrinkage target (Zhang et al., 2020); or the Combination Elastic Net (Rapach & Zhou, 2020). However, in many practical applications, the simple average of candidate forecasts is more robust than more sophisticated combination approaches (Qian et al., 2019), a phenomenon known as the *forecast combination puzzle*. A theoretical explanation for the latter is given in Claeskens et al. (2016) as well as the warning that " there is no guarantee that the 'optimal' forecast combination will be better than the equal-weight case, or even improve on the original forecasts".

3 Methodology and materials

In this section, we introduce the underlying financial model and the corresponding nonparametric predictive long-term regressions. We follow the approach of Scholz et al. (2015) and focus on (nonlinear) relationships between stock returns in excess of different benchmarks and a set of predictor variables. We aim to compare individual models with several combination approaches in terms of their in-sample and out-of-sample predictability over the horizon of 1 year. We consider the four benchmarks introduced in Kyriakou et al. (2021a): the short- and the long-term interest rate, the earnings-by-price ratio, and the inflation rate.

3.1 Predictive framework

Let D_t denote the (nominal) dividends paid during year t and P_t the (nominal) stock price at the end of year t. We consider stock returns $S_t = (P_t + D_t)/P_{t-1}$ in excess (log-scale) of a given reference rate or benchmark $B_{t-1}^{(A)}$:

$$Y_t^{(A)} = \ln \frac{S_t}{B_{t-1}^{(A)}}, \tag{2}$$

where $A \in \{R, L, E, C\}$ with, respectively,

$$B_t^{(R)} = 1 + \frac{R_t}{100}, \quad B_t^{(L)} = 1 + \frac{L_t}{100}, \quad B_t^{(E)} = 1 + \frac{E_t}{P_t}, \quad B_t^{(C)} = \frac{CPI_t}{CPI_{t-1}},$$

R_t is the short-term interest rate, L_t the long-term interest rate, E_t the earnings accruing to the index in year t, and CPI_t the consumer price index for year t. The predictive nonparametric regression model for the 1-year excess stock returns $Y_t^{(A)}$ is then given by

$$Y_t^{(A)} = m(X_{t-1}^{(A)}) + \xi_t, \tag{3}$$

where

$$m(x^{(A)}) = \mathbb{E}(Y^{(A)} | X^{(A)} = x^{(A)}), \; x^{(A)} \in \mathbb{R}^q, \tag{4}$$

is the unknown conditional mean-function which is estimated with the local-linear smoother. The error-terms ξ_t in Eq. (3) form a martingale difference process and are serially uncorrelated zero-mean random variables of an unknown conditionally heteroscedastic form $\sigma(x)$.

Our individual predictive models use (subsets of) popular time-lagged predictive variables: the dividend-by-price ratio $d_{t-1} = D_{t-1}/P_{t-1}$; the earnings-by-price ratio $e_{t-1} = E_{t-1}/P_{t-1}$; the short-term interest rate $r_{t-1} = R_{t-1}/100$; the long-term interest rate $l_{t-1} = L_{t-1}/100$; the inflation rate $\pi_{t-1} = (CPI_{t-1} - CPI_{t-2})/CPI_{t-2}$; the term spread $s_{t-1} = l_{t-1} - r_{t-1}$; and the excess stock return $Y_{t-1}^{(A)}$. Note that we apply both the single benchmarking approach (Kyriakou et al., 2021a), where only the dependent variable in Eq. (2) is transformed with the benchmark, and the double benchmarking approach, where also the predictive variables are transformed according to

$$X_{t-1}^{(A)} = \begin{cases} \frac{1+X_{t-1}}{B_{t-1}^{(A)}}, & X \in \{d, e, r, l, \pi\} \\ \frac{s_{t-1}}{B_{t-1}^{(A)}} = \frac{l_{t-1}-r_{t-1}}{B_{t-1}^{(A)}} & \\ Y_{t-1}^{(A)} & \end{cases}, \quad A \in \{R, L, E, C\}. \tag{5}$$

The double benchmarking approach can be seen as a simple way of reducing dimensionality. It allows to import more structure in the estimation process which can help to reduce or circumvent problems caused by the curse of dimensionality. Remember that we apply our methods to annual data, that is, we use sparsely distributed observations in higher dimensions which limits the complexity of the fitted models.

3.2 Estimation and evaluation procedure

In the empirical part, we estimate the unknown conditional mean function m of Eq. (3) with the local-linear smoother which is based on the following minimization problem

$$\min_{a,b} \sum_{t=1}^{T} \left(Y_t^{(A)} - a - \left(X_t^{(A)} - x^{(A)} \right)^{\top} b \right)^2 K_h \left(X_t^{(A)} - x^{(A)} \right), \tag{6}$$

where K_h denotes some kernel function, for example, the standard product kernel $K_h \left(X_t^{(A)} - x^{(A)} \right) = \prod_{s=1}^{q} \frac{1}{h_s} k \left(\frac{X_{t,s}^{(A)} - x_s^{(A)}}{h_s} \right)$ which depends on a set of bandwidths $h = (h_1, \ldots, h_q)$ and the kernels k of order ν. The latter are univariate symmetric functions satisfying standard assumptions: $\int k(u)du = 1$, $\int u^l k(u)du = 0$ $(l = 1, \ldots, \nu - 1)$, and $\int u^{\nu} k(u)du =: \kappa_{\nu} > 0$. $X_{t,s}^{(A)}$ denotes the sth component of $X_t^{(A)}$, $s = 1, \ldots, q$. The solution $\hat{a} = \hat{a}(x^{(A)})$ of (6) is a consistent estimator of $m(x^{(A)})$ which depends on the bandwidths h. For a discussion of properties and references for proofs, see, for example, Section 3.1 in Kyriakou et al. (2021a).

For the choice of the smoothing parameters h, we apply the local linear cross-validation approach and select those bandwidths which minimize

$$CV(h) = \min_{h} \sum_{t=1}^{T} \left(Y_t^{(A)} - \hat{m}_{-t,h} \right)^2, \tag{7}$$

where T is the number of observations in the estimation sample and $\hat{m}_{-t,h}$ is the leave-k-out estimator for the conditional mean function. It is computed by removing k observations around the tth time point and depends on the horizon of the prediction. Here we focus on the 1-year horizon and use the classical leave-one-out estimator.

Based on the cross-validation criterion in Eq. (7), we introduce next our validation measure used for in-sample model selection. It is a generalization of the validated R^2 (Nielsen & Sperlich, 2003) and is defined as

$$R_V^2 = 1 - \frac{\sum_{t=1}^{T} \left(Y_t^{(A)} - \hat{m}_{-t,h} \right)^2}{\sum_{t=1}^{T} \left(Y_t^{(A)} - \bar{Y}_{-t}^{(A)} \right)^2}, \tag{8}$$

where leave-k-out estimators ($\hat{m}_{-t,h}$ and $\bar{Y}_{-t}^{(A)}$) are used for the conditional mean function m and for the unconditional (historical) mean of $Y_t^{(A)}$, respectively. The R_V^2 measures the predictive power of a given model compared to the cross-validated historical mean. A positive R_V^2 implies that the predictor-based regression model (3) outperforms the corresponding historical average excess stock return over T years. Thus, we use the R_V^2 to rank all possible candidate models and prefer the one with the largest value. We can use the R_V^2 also for bandwidth selection as maximizing the R_V^2 in Eq. (8) is equivalent to minimizing the cross-validation criterion in Eq. (7). Note further that we apply the R_V^2 also to the linear counterparts of the regression model (3). In this case, just replace $\hat{m}_{-t,h}$ by the linear predictor based on the leave-k-out OLS-estimate $\hat{\beta}_{-t}$.

For out-of-sample evaluation, we use the last τ observations in our records to calculate the classical out-of-sample R^2 (Campbell & Thompson, 2008) which is defined as

$$R^2_{oos} = 1 - \frac{\sum_{t=T+1}^{T+\tau} \left(Y_t^{(A)} - \hat{m}_t\right)^2}{\sum_{t=T+1}^{T+\tau} \left(Y_t^{(A)} - \bar{Y}_t^{(A)}\right)^2}, \tag{9}$$

where $\hat{m}_t$ is the fitted value from the predictive regression estimated through period T (the last observation in the estimation sample) and evaluated at $X_{t-1}^{(A)}$ ($t = T+1, \ldots, T+\tau$), and $\bar{Y}_t^{(A)}$ is the historical average return through period $t-1$. In other words, we use the estimation sample ($t = 1, \ldots, T$) to fix the model by choosing corresponding bandwidths and evaluate through period $t-1$ in the left out sample. A positive R^2_{oos} indicates that the predictive regression has a lower average mean squared prediction error than the historical average return. As Campbell and Thompson (2008) point out, the historical average has an advantage over predictive regressions because it is based on more observations and more recently available information.

3.3 Forecast combinations

It is well documented in the literature that the combination of M individual forecasts $\hat{Y}_{t+1}^{(A),m}$ (with $m = 1, \ldots, M$), defined as

$$\hat{Y}_{t+1}^{comb} = w_1 \hat{Y}_{t+1}^{(A),1} + \ldots + w_M \hat{Y}_{t+1}^{(A),M}, \tag{10}$$

may perform better (in terms of higher out-of-sample predictability) than the individual predictions itself (Bates & Granger, 1969; Granger & Ramanathan, 1984; Rapach et al., 2010). A popular choice is, for example, the simple average of the M different predictors:

$$\hat{Y}_{t+1}^{av} = \frac{1}{M} \sum_{m=1}^{M} \hat{Y}_{t+1}^{(A),m}. \tag{11}$$

Each individual forecast gets the same weight $w_m = 1/M$ which shrinks, in case of a multivariate linear predictive model, the estimated (and probably biased) coefficients by the factor $1/M$ and reduces the role of multicollinearity when highly correlated predictors are used (Rapach et al., 2010). The simple average (11) allows to incorporate information of a large number of plausible predictors and helps to prevent from in-sample over-fitting (Rapach & Zhou, 2020). However, equal weights can be sub-optimal as one usually wants to give more weight to those forecasts with errors of lower variance (Diebold & Shin, 2019). In addition, when a large number of potential predictors is available, the redundant ones should be excluded, that is, have a weight of zero. Thus, several ML techniques have been applied to select and weight the relevant predictors in Eq. (10). Popular regularization methods set some weights to zero and shrink the remaining weights to zero [the 'classical' Lasso (Tibshirani, 1996), the 'adaptive' Lasso (Zou, 2006), the Ridge Regression (Hoerl & Kennard, 1970), or the Elastic Net (ENet) (Zou & Hastie, 2005)) or toward equality [the 'egalitarian' Lasso, the 'egalitarian' Ridge (Diebold & Shin, 2019), or the 'combination' Elastic Net (cENet) (Rapach & Zhou, 2020)].

The underlying penalization problem for the forecast combination methods used in this paper can be summarized as follows:

$$\hat{w} = \arg\min_w \left[\sum_{t=1}^{T} \left(Y_t^{(A)} - \sum_{m=1}^{M} w_m \hat{Y}_t^{(A),m} \right)^2 + \lambda \sum_{m=1}^{M} \left\{ \alpha |w_m| + (1-\alpha) w_m^2 \right\} \right], \quad (12)$$

that is, the Lasso ($\alpha = 1$), the Ridge ($\alpha = 0$), and the ENet ($\alpha \in (0, 1)$) with w_m ($m = 1, \ldots, M$) restricted to be non-negative. In addition, we consider their 'egalitarian' versions (eLasso, eRidge, and eENet) using a two-step procedure (Diebold & Shin, 2019): Solving the standard problem (12), we (i) select the l important forecasts of the full set of M potential candidates, that is, the $M - l$ forecasts with weight zero are excluded; and (ii) shrink their combining weights towards equality, that is, toward $1/l$. Recently, Rapach and Zhou (2020) proposed a further refinement, the so called 'combination' ENet (cENet). They split the estimation sample into two parts: an initial in-sample period and an 'holdout' out-of-sample period; and apply the eENet only on the latter instead on all available observations. Note further that we use the multivariate regression approach introduced in Sect. 3.1. Therefore, we also account for model complexity measured by the number of included predictor variables and combine the different forecasts based on complete subset regressions with dimensionality $k \in \{1, 2, 3\}$ (Elliott et al., 2013).

This leads in total to 32 different ways of combining the individual forecasts: applying the set of methods consisting of Lasso, Ridge, ENet, eLasso, eRidge, eENet, cENet, and simple average to all available potential forecasts or restricting to the k-dimensional ones with $k \in \{1, 2, 3\}$.

3.4 The data

In the empirical part of this paper, we apply the methods described in Sects. 3.2 and 3.3 to annual US stock market data over the period 1872 to 2022. We use a revised and updated version of the series described in Shiller's Chapter 26 (Shiller, 1989) which consist of the Standard and Poor's (S&P) Composite Stock Price Index, dividends and earnings accruing to the index, a 1-year interest rate, a long government bond yield, and the consumer price index.[1] Note that we had to replace the original risk-free rate series (which was discontinued in 2013) by an annual yield based on the 6-month Treasury-bill rate,[2] secondary market. This new series is only available from 1958 onwards. Therefore, we regressed the Treasury-bill rate on the original commercial paper rate from Shiller's data and instrumented the risk-free rate from 1872 to 1957 with corresponding predicted values. For more details, see, for example, Kyriakou et al. (2020) or Mammen et al. (2019). Table 1 summarizes the available variables with their basic descriptive statistics for both, the in-sample part of the data used for estimation and the out-of-sample part of the data used for evaluation of predictability. It is evident that most of the variables have a much larger mean and standard deviation in the left-out part. However, we focus on the predictability of excess stock returns which are very similar in both parts. Figure 1 exemplarily shows them in excess of the risk-free rate with the out-of-sample period highlighted in red. Note that large positive returns have been realized with higher probability in the in-sample part of the data.

[1] http://www.econ.yale.edu/~shiller/data.htm.

[2] https://fred.stlouisfed.org/series/TB6MS.

Table 1 US market data (1872–2022)

	Max	Min	Mean	Sd	Skew	Exc. kurt
A: In-sample part of data (1872–1962)						
S&P stock price index	69.07	3.25	13.16	13.24	2.47	5.81
Dividend accruing to index	2.13	0.18	0.63	0.49	1.51	1.29
Earnings accruing to index	3.67	0.16	1.04	0.93	1.52	1.13
Short-term interest rate	7.46	0.55	3.46	1.69	− 0.08	− 0.57
Long-term interest rate	5.58	1.95	3.56	0.82	0.29	− 0.28
Consumer price index	30.00	6.47	14.30	6.74	0.89	− 0.32
B: Out-of-sample part of data (1963–2022)						
S&P stock price index	4573.82	65.06	863.33	999.28	1.65	2.64
Dividend accruing to index	60.4	2.28	17.16	15.96	1.30	0.73
Earnings accruing to index	197.84	4.02	40.70	41.55	1.49	2.05
Short-term interest rate	14.93	0.07	4.68	3.29	0.56	0.33
Long-term interest rate	14.59	1.08	5.87	2.97	0.63	0.03
Consumer price index	281.15	30.4	137.92	76.5	0.07	− 1.34

The S&P 500 Stock Composite Index in its current form was introduced in March 1957. Relevant series prior to this date have been re-calculated or taken from several historical sources. For more details, see, Chapter 26 in Shiller (1989)

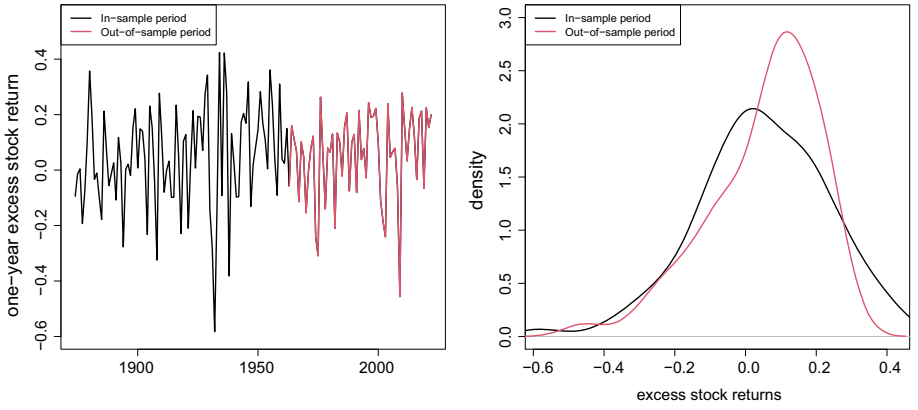

Fig. 1 Stock returns in excess of the risk-free rate. In-sample part (black), out-of-sample part (red). Left: Time-series plot, Right: Density estimates. Period: 1872–2022. Data: annual S&P 500. (Color figure online)

4 Results and discussion

In this section, we present and discuss the results of the empirical application. For ease of presentation, we focus to the most important benchmark models of Kyriakou et al. (2021a), the short-term interest rate (single benchmarking) and the inflation rate (double benchmarking). Results for the other benchmarks (single and double benchmarking) are available upon request. Note that the short-term interest rate benchmark directly corresponds to the classical prediction of the risk premium (over a risk-free investment) and the inflation rate benchmark refers to the forecast of real returns as led by Merton (2014).

One primary goal of this study is to compare the in-sample predictive power of several methods with the corresponding out-of-sample performance. For this reason, we split the annual US stock market data into two parts: (i) an in-sample period (1872–1962) used for (smoothing) parameter estimation and in-sample validation and (ii) an out-of-sample period (1963–2022) of 60 years used for 1-year-ahead prediction and out-of-evaluation. We estimated the nonparametric models with a local-linear kernel smoother using the quartic (product) kernel. The smoothing parameters (bandwidths) were chosen by leave-one-out cross-validation, that is, by maximizing the in-sample performance measure R^2_V introduced in Sect. 3.2. In other words, the in-sample period is just used to fix the smoothness of the underlying conditional mean function. The prediction itself is then based on most recent (time-lagged) information. The corresponding linear models were estimated with ordinary least squares (OLS).

4.1 Prediction of the risk-premium

Numerous academic research articles rely on macroeconomic variables to forecast the U.S. equity risk premium. We follow this road and present to begin with in Table 2 the comparison of in-sample predictive power (measured by the R^2_V) and out-of-sample performance (measured by the R^2_{oos}) for several individual models. Based on the in-sample measure, the best five nonparametric models ($\{sp\}$, $\{r\}$, $\{r, sp\}$, $\{l, sp\}$, $\{r, l\}$) have a R^2_V in the range of 8.8–6.9% and are one- or two-dimensional models with three of them including the term spread as covariate. However, only two of those models ($\{sp\}$,$\{l, sp\}$) perform convincingly out-of-sample and are under the top five predictive models ($\{l, sp\}$, $\{Y, d, sp\}$, $\{e, inf, sp\}$, $\{l, inf, sp\}$, $\{sp\}$) demonstrating a R^2_{oos} in the range of 15.5–9.1%. Note that the term-spread is included in all of these models and that now also some of the three-dimensional models perform reasonably out-of-sample. Nevertheless, most three-dimensional models cannot beat the historical mean over the considered 60 year out-of-sample period. For the linear models, we find a similar set of five best performing (in-sample) models ($\{sp\}$, $\{r\}$, $\{r, l\}$, $\{r, sp\}$, $\{l, sp\}$) with a R^2_V in the lower range of 7.9–6.6%. However, only three of those models can beat the historical mean out-of-sample ($\{e, sp\}$, $\{r, sp\}$, $\{l, sp\}$ with a R^2_{oos} between 4.2% and 7.0%). The five best performing predictive models ($\{d, r, l\}$, $\{d, l, sp\}$, $\{e, r, l\}$, $\{e, r, sp\}$, $\{e, l, sp\}$) are all three-dimensional with R^2_{oos} in the range of 13.2–9.3%. Here, the combination of earnings and spread together with an additional variable gives the most promising results.

In a next step, we consider the correlation between the individual forecasts. Figure 2 displays the correlation matrix of forecasts from all one- and two-dimensional nonparametric models for both the in-sample (left) and out-of sample predictions (right). The 'ideal' or best individual predictor would be (highly) positively correlated with the excess stock returns in the out-of-sample period. To improve over such a forecast, the (linear) combination of different individual predictors must be (i) positively correlated to the former and (ii) would be composed of positively correlated candidates. When considering forecasts from the two-dimensional models, the left-hand side of Fig. 2 shows for most of them a (high) positive correlation. Thus, one would expect a large potential for improvements in predictive power using forecast combinations when a strong selection (shrinkage) of only a few predictive candidates based on the in-sample information occurs. However, for the corresponding out-of-sample predictions the correlations are less pronounced and for some even negative. Using now the weights fixed in the estimation sample will not necessarily lead to an improved out-of-sample performance. For a theoretical analysis on factors that determine the advantages from

Table 2 Comparison of predictive power: in-sample (measured by the R^2_V) versus out-of-sample (measured by the R^2_{oos})

Model		Non-par		Linear		Model		Non-par		Linear	
Nr.	X	R^2_V	R^2_{oos}	R^2_V	R^2_{oos}	Nr.	X	R^2_V	R^2_{oos}	R^2_V	R^2_{oos}
1.	Y	− 0.020	− 0.004	− 0.018	− 0.004	33.	Y, d, sp	0.037	**0.112**	0.041	0.093
2.	d	− 0.016	− 0.020	− 0.016	− 0.008	34.	Y, e, r	0.028	− 0.590	0.035	− 0.223
3.	e	− 0.013	− 0.021	− 0.011	− 0.017	35.	Y, e, l	− 0.016	− 1.287	− 0.008	− 0.997
4.	r	**0.084**	− 1.639	**0.073**	− 0.304	36.	Y, e, inf	− 0.034	− 0.086	− 0.025	− 0.081
5.	l	0.045	0.009	0.022	− 1.099	37.	Y, e, sp	0.034	0.064	0.039	0.045
6.	inf	− 0.022	0.011	− 0.020	0.011	38.	Y, r, l	0.032	− 0.355	0.036	0.055
7.	sp	**0.088**	**0.091**	**0.079**	− 0.018	39.	Y, r, inf	0.021	− 0.686	0.024	− 0.325
8.	Y, d	− 0.019	− 0.197	− 0.014	− 0.178	40.	Y, r, sp	0.041	− 1.067	0.036	0.055
9.	Y, e	− 0.031	− 0.039	− 0.026	− 0.036	41.	Y, l, inf	− 0.027	− 0.097	− 0.021	− 1.091
10.	Y, r	0.041	− 0.648	0.042	− 0.291	42.	Y, l, sp	0.034	0.003	0.036	0.055
11.	Y, l	− 0.010	− 0.057	− 0.004	− 0.957	43.	Y, inf, sp	0.030	0.015	0.033	− 0.009
12.	Y, inf	− 0.037	− 0.004	− 0.034	− 0.005	44.	d, e, r	0.026	− 0.663	0.034	− 0.241
13.	Y, sp	0.050	0.047	0.049	− 0.021	45.	d, e, l	− 0.025	− 1.612	− 0.017	− 1.404
14.	d, e	− 0.050	− 0.024	− 0.043	− 0.014	46.	d, e, inf	− 0.058	− 0.010	− 0.048	− 0.005
15.	d, r	0.061	− 0.250	0.059	− 0.334	47.	d, e, sp	0.031	0.021	0.038	− 0.027
16.	d, l	0.010	− 0.229	0.013	− 1.624	48.	d, r, l	0.048	− 0.165	0.050	**0.093**

Table 2 continued

Model		Non-par		Linear		Model		Non-par		Linear	
Nr.	X	R^2_V	R^2_{oos}	R^2_V	R^2_{oos}	Nr.	X	R^2_V	R^2_{oos}	R^2_V	R^2_{oos}
17.	d, inf	− 0.041	− 0.016	− 0.037	− 0.005	49.	d, r, inf	0.032	− 0.775	0.037	− 0.356
18.	d, sp	0.061	0.062	0.061	0.010	50.	d, r, sp	0.052	− 0.136	0.050	0.093
19.	e, r	0.064	− 0.575	0.065	− 0.243	51.	d, l, inf	− 0.016	− 0.266	− 0.009	− 1.755
20.	e, l	0.014	− 0.124	0.016	− 1.174	52.	d, l, sp	0.049	0.041	0.050	**0.093**
21.	e, inf	− 0.029	− 0.035	− 0.024	− 0.029	53.	d, inf, sp	0.036	0.070	0.040	0.015
22.	e, sp	0.065	0.071	**0.066**	0.042	54.	e, r, l	0.056	− 0.202	0.055	**0.132**
23.	r, l	**0.069**	− 0.420	0.066	0.070	55.	e, r, inf	0.043	− 0.612	0.052	− 0.289
24.	r, inf	0.053	− 0.740	0.052	− 0.329	56.	e, r, sp	0.057	− 0.046	0.055	**0.132**
25.	r, sp	**0.077**	− 0.538	**0.066**	0.070	57.	e, l, inf	0.000	− 0.235	0.008	− 1.519
26.	l, inf	− 0.001	− 0.105	0.002	− 1.242	58.	e, l, sp	0.056	− 0.700	0.055	**0.132**
27.	l, sp	**0.073**	**0.155**	**0.066**	0.070	59.	e, inf, sp	0.044	**0.095**	0.050	0.075
28.	inf, sp	0.058	0.019	0.058	− 0.010	60.	r, l, inf	0.042	− 0.110	0.045	0.069
29.	Y, d, e	− 0.058	− 0.177	− 0.047	− 0.155	61.	r, l, sp	–	–	–	–
30.	Y, d, r	0.039	− 0.872	0.042	− 0.411	62.	r, inf, sp	0.045	0.024	0.045	0.069
31.	Y, d, l	0.002	− 1.389	0.009	− 1.854	63.	l, inf, sp	0.044	**0.093**	0.045	0.069
32.	Y, d, inf	− 0.038	− 0.229	− 0.030	− 0.215						

One-year stock returns in excess of the short-term interest rate $Y_t^{(R)}$ modelled with the nonparametric smoother and its linear counterpart (the single benchmarking approach) for several combinations of predictive variables X. In-sample period: 1872–1962, out-of-sample period: 1963–2022. Largest five results (column-wise) highlighted in bold

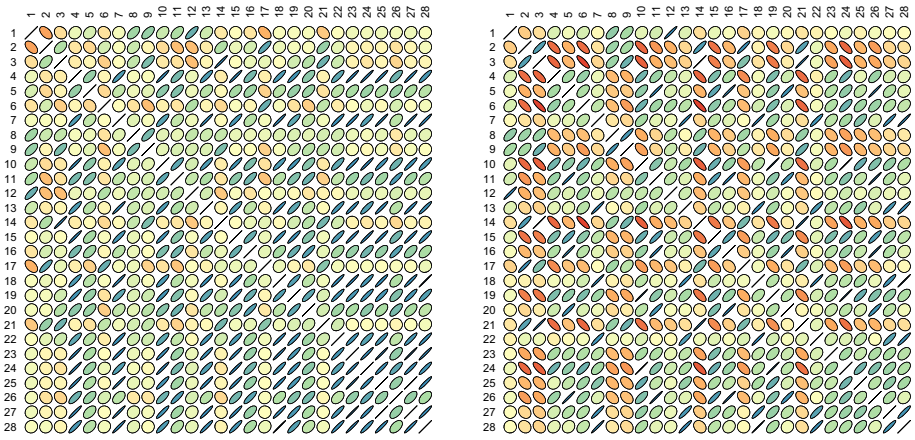

Fig. 2 Correlations of predictions for stock returns in excess of the risk-free rate (for nonlinear models of one or two predictive variables). Left: In-sample, Right: Out-of-sample. Period: 1872–2022. Data: annual S&P 500

combining forecasts including a discussion on their correlation can be found, for example, in Timmermann (2006). Note that there are only a few studies which directly account for the possibility of correlation between forecasts (Guerrero & Pena, 2003).

As described in Sect. 3.3, forecast combinations are a popular method for further improving the forecast quality. Table 3 summarizes 32 different versions of such combinations making use of the individual forecasts shown in Table 2. When using all 62 different nonparametric forecasts, the ENet (25.1%), the Ridge (22.3%), the Lasso (21.6%), and the eLasso (8.8%) improve in-sample over the individual models. However, none of those combinations produces forecasts that can beat the historical mean out-of-sample, that is, having any predictive power. This finding shows that those methods are prone to in-sample over-fitting when too many candidate models are available (even when these methods are validated against the mean). The situation is different, when the possible candidates are restricted to be one-dimensional. Now, the Enet, the Lasso, and the eLasso have both higher in-sample and higher out-of-sample power than individual models (R^2_V =13.1%, 11.4%, 9.2%, and R^2_{oos} =17.1%, 17.0%, 16.1%). Note that all of those combine individual forecasts based on the term-spread and the long-term interest rate. In terms of out-of-sample improvements, restrictions to two- or three-dimensional individual forecasts are less successful strategies. For example, in the two-dimensional case, the Enet (R^2_V = 11.0%) selects the forecasts of the following four individual models: $\{Y, sp\}$, $\{d, r\}$, $\{e, r\}$, and $\{e, inf\}$, which are highly correlated in-sample. The combined forecast improves over the individual ones in-sample but is far away from the out-of-sample predictive power of the best individual model. For the linear counterpart, only a few of the forecast combination methods are able to increase the in-sample performance. For example, the Ridge and the Enet restricted to three-dimensional individual models show R^2_V values of 9.5% and 8.0%. However, only one of the 32 ways of combining individual forecasts was able to improve out-of-sample over the best individual three-dimensional model (eLasso with R_{oos} = 13.6). Note also that in-sample over-fitting is not such an issue in the linear case because the higher-dimensional models do not include interaction terms. This is also the reason for having very similar results, when accounting for model complexity (that is, similar R^2_{oos} values in all the panels of Table 3). The nonlinear and linear case have in common that the recently proposed refinement of the elastic net, the cEnet, was hardly able

Table 3 Comparison of predictive power: in-sample (measured by the R^2_V) versus out-of-sample (measured by the R^2_{oos})

Forecast combination			Non-par		Linear	
Nr.	Dim.	type	R^2_V	R^2_{oos}	R^2_V	R^2_{oos}
64.	All	Lasso	**0.216**	− 1.013	− 0.001	0.112
65.	All	Ridge	**0.223**	− 1.116	0.072	0.053
66.	All	Enet	**0.251**	− 1.321	**0.089**	0.124
67.	All	eLasso	**0.088**	− 0.282	0.054	0.109
68.	All	eRidge	0.009	− 0.027	0.053	0.043
69.	All	eEnet	0.035	− 0.035	0.056	0.107
70.	All	cENet	0.022	0.000	0.054	0.095
71.	All	Average	0.054	0.072	0.053	0.043
72.	1D	Lasso	**0.114**	**0.170**	0.062	0.095
73.	1D	Ridge	**0.117**	− 0.067	0.063	0.029
74.	1D	Enet	**0.131**	**0.171**	0.075	0.101
75.	1D	eLasso	**0.092**	**0.161**	**0.080**	0.053
76.	1D	eRidge	0.051	0.025	0.041	0.041
77.	1D	eEnet	**0.091**	0.146	0.078	0.035
78.	1D	cENet	0.033	0.046	0.022	0.000
79.	1D	Average	0.051	0.025	0.041	0.041
80.	2D	Lasso	0.067	0.099	0.024	0.116
81.	2D	Ridge	0.087	0.090	0.067	0.046
82.	2D	Enet	**0.110**	0.116	0.075	0.123
83.	2D	eLasso	0.065	0.072	0.068	0.121
84.	2D	eRidge	0.055	0.087	0.054	0.032
85.	2D	eEnet	0.034	0.045	0.064	0.112
86.	2D	cENet	0.022	0.000	0.044	0.098
87.	2D	Average	0.055	0.087	0.054	0.032
88.	3D	Lasso	**0.254**	− 1.124	0.057	0.128
89.	3D	Ridge	**0.217**	− 1.496	**0.095**	0.058
90.	3D	Enet	**0.258**	− 1.103	**0.080**	0.125
91.	3D	eLasso	0.041	− 1.067	0.054	**0.136**
92.	3D	eRidge	0.021	− 0.095	0.050	0.043
93.	3D	eEnet	0.041	− 1.067	0.053	0.123
94.	3D	cENet	0.045	0.098	0.052	− 0.289
95.	3D	Average	0.049	0.060	0.050	0.043

One-year stock returns in excess of the short-term interest rate $Y_t^{(R)}$ modelled by forecast combinations based on individual forecasts of models from Table 2 (the single benchmarking approach) applying the Lasso, the Ridge, the Enet, the eLasso, the eRidge, the eEnet, the cENet, and the simple average to all possible models or only to the k-dimensional ones with $k \in \{1, 2, 3\}$. In-sample period: 1872–1962, out-of-sample period: 1963–2022. Improvements (column-wise) compared to largest measure in Table 2 highlighted in bold

to beat the historical mean at all. For the linear models, it was even the only of the eight different ways of combining individual forecasts that produced negative R^2_{oos} values.

Now, we address the question of how the models best performing out-of-sample behave during recessions and economic expansions. For this purpose, we calculate the out-of-sample mean squared error during the aforementioned sub-samples based on the US business cycle expansion and contraction data provided by the NBER[3] and for the full period. Note that in the 60-year out-of-sample period only 8 years have been classified as recessions years (that is, with more than 6 months of a recession). Table 4 shows a comparison of these out-of-sample measures for the individual models, while Table 5 focusses on the forecast combination methods. It is evident that the (nonparametric and linear) models with the smallest out-of-sample mean squared error over the full period (and thus largest R^2_{oos} values) belong to the best performing models during economic expansions. Note that such models perform only slightly better than the historical mean during the recessions. There are several models which perform reasonably well during the recessions (for example, $\{e, l, inf\}$ or $\{d, l, inf\}$). However, they have in common not to be able to beat the historical mean during economic expansions. A similar conclusion can be drawn for the forecast combination methods. Only a few of them can improve over the best individual models during the full period and the expansions, while non of those methods improves during the recessions.

We finish the empirical analysis for the risk premium by checking the robustness of the considered models over time. For this purpose, we increased the in-sample period stepwise from 89 to 124 years (and reduced the out-of-sample evaluation period correspondingly from 60 to 35 years). Figure 3 shows the R^2_V (left) and the R^2_{oos} (right) for models with the largest out-of-sample R^2 (we show the best three nonparametric and the best three linear models, resp.). Note that for the nonparametric models, only individual models give reasonable results over time. The best three are: $\{sp\}$, $\{e, sp\}$, and $\{d, inf, sp\}$. However, the most forecast combination models suffered from negative R_{oos} during the out-of-sample periods 1975–2022 or 1980–2022. For the linear models, the situation is quite different: only three individual models but most forecast combination models performed steadily over time. The best three in this case are: the Lasso and the Enet over all individual models, and the Lasso over all 3-dim. models. Figure 3 shows as well that (i) the in-sample performance of the displayed models increases steadily over time and (ii) the out-of-sample performance for the first half of the considered period remains stable but reduces sharply at their end. A possible explanation could be the fact that when the out-of-sample period gets shorter and shorter it is highly dominated by the large negative returns during the Great Recession which was caused by the Global Financial Crisis (compare also Fig. 1). To summarize, the model which performed best in terms of a high and stable R^2_{oos} was the nonparametric model based on the term-spread as covariate.

4.2 Prediction of real returns

Real-income protection is one of the main aspects in long-term pension planning (Merton, 2014; Gerrard et al., 2018, 2019). Therefore, the underlying financial model used when optimizing the investment asset allocation for the long term should reflect these needs in real terms. We apply here the double benchmarking approach of Kyriakou et al. (2021a) with the inflation as the reference rate, that is, all variables are measured net of inflation. Note that inflation itself cannot be included as a covariate because it is transformed to a constant. Therefore, only 40 different models are possible under the inflation benchmark

[3] https://fred.stlouisfed.org/series/USREC.

Table 4 Comparison of predictive power: out-of-sample mean squared error for the full sample period (full), during recessions (rec) and expansions (exp)

Model		Non-par			Linear			Model		Non-par			Linear		
Nr.	*X*	Full	Rec	Exp	Full	Rec	Exp	Nr.	*X*	Full	Rec	Exp	Full	Rec	Exp
1.	Y	0.024	0.072	0.017	0.024	0.072	0.017	33.	Y, d, sp	**0.021**	0.053	0.016	0.022	0.053	0.017
2.	d	0.024	0.065	0.018	0.024	0.067	0.017	34.	Y, e, r	0.038	0.074	0.033	0.029	**0.045**	0.027
3.	e	0.024	0.071	0.017	0.024	0.071	0.017	35.	Y, e, l	0.055	0.046	0.056	0.048	**0.041**	0.049
4.	r	0.063	0.067	0.063	0.031	0.055	0.028	36.	Y, e, inf	0.026	0.055	0.021	0.026	0.056	0.021
5.	l	0.024	0.052	0.019	0.050	0.053	0.050	37.	Y, e, sp	0.022	0.061	0.016	0.023	0.062	0.017
6.	inf	0.024	0.071	0.016	0.024	0.072	**0.016**	38.	Y, r, l	0.032	0.116	0.020	0.023	0.060	0.017
7.	sp	**0.022**	0.067	**0.015**	0.024	0.069	0.017	39.	Y, r, inf	0.040	0.081	0.034	0.032	0.052	0.029
8.	Y, d	0.029	0.048	0.026	0.028	0.048	0.025	40.	Y, r, sp	0.049	0.064	0.047	0.023	0.060	0.017
9.	Y, e	0.025	0.067	0.018	0.025	0.067	0.018	41.	Y, l, inf	0.026	0.047	0.023	0.050	0.049	0.050
10.	Y, r	0.039	0.082	0.033	0.031	0.053	0.028	42.	Y, l, sp	0.024	0.065	0.017	0.023	0.060	0.017
11.	Y, l	0.025	0.051	0.021	0.047	0.049	0.046	43.	Y, inf, sp	0.024	0.065	0.017	0.024	0.065	0.018
12.	Y, inf	0.024	0.067	0.017	0.024	0.067	0.017	44.	d, e, r	0.040	0.083	0.033	0.030	0.048	0.027
13.	Y, sp	0.023	0.066	0.016	0.024	0.067	0.018	45.	d, e, l	0.063	0.052	0.064	0.058	0.046	0.059
14.	d, e	0.024	0.070	0.017	0.024	0.072	0.017	46.	d, e, inf	0.024	0.068	0.017	0.024	0.070	0.017
15.	d, r	0.030	0.047	0.027	0.032	0.050	0.029	47.	d, e, sp	0.023	0.070	0.016	0.025	0.071	0.017
16.	d, l	0.029	**0.042**	0.027	0.063	0.049	0.065	48.	d, r, l	0.028	0.097	0.017	**0.022**	0.055	0.017
17.	d, inf	0.024	0.062	0.018	0.024	0.065	0.018	49.	d, r, inf	0.042	0.084	0.036	0.032	0.050	0.030
18.	d, sp	0.022	0.065	0.016	0.024	0.067	0.017	50.	d, r, sp	0.027	0.056	0.023	0.022	0.055	0.017
19.	e, r	0.038	0.074	0.032	0.030	0.048	0.027	51.	d, l, inf	0.030	**0.040**	0.029	0.066	0.051	0.068
20.	e, l	0.027	0.046	0.024	0.052	0.046	0.053	52.	d, l, sp	0.023	0.063	0.017	**0.022**	0.055	0.017
21.	e, inf	0.025	0.063	0.019	0.025	0.064	0.019	53.	d, inf, sp	0.022	0.064	0.016	0.024	0.066	0.017

Table 4 continued

Model		Non-par			Linear			Model		Non-par			Linear		
Nr.	X	Full	Rec	Exp	Full	Rec	Exp	Nr.	X	Full	Rec	Exp	Full	Rec	Exp
22.	e, sp	0.022	0.064	**0.016**	0.023	0.065	0.016	54.	e, r, l	0.029	0.099	0.018	**0.021**	0.053	**0.016**
23.	r, l	0.034	0.075	0.028	0.022	0.060	0.016	55.	e, r, inf	0.039	0.070	0.034	0.031	0.046	0.029
24.	r, inf	0.042	0.084	0.035	0.032	0.055	0.028	56.	e, r, sp	0.025	0.051	0.021	**0.021**	0.053	**0.016**
25.	r, sp	0.037	0.143	0.020	0.022	0.060	0.016	57.	e, l, inf	0.030	**0.039**	0.028	0.060	0.046	0.062
26.	l, inf	0.026	0.048	0.023	0.054	0.054	0.054	58.	e, l, sp	0.041	0.059	0.038	**0.021**	0.053	**0.016**
27.	l, sp	**0.020**	0.062	**0.014**	0.022	0.060	0.016	59.	e, inf, sp	**0.022**	0.060	**0.016**	0.022	0.061	**0.016**
28.	inf, sp	0.023	0.068	0.017	0.024	0.068	0.017	60.	r, l, inf	0.027	0.056	0.022	0.022	0.059	0.017
29.	Y, d, e	0.028	0.050	0.025	0.028	0.051	0.024	61.	r, l, sp	–	–	–	–	–	–
30.	Y, d, r	0.045	0.076	0.040	0.034	**0.038**	0.033	62.	r, inf, sp	0.023	0.059	0.018	0.022	0.059	0.017
31.	Y, d, l	0.057	**0.042**	0.060	0.068	**0.036**	0.073	63.	l, inf, sp	**0.022**	0.061	**0.016**	0.022	0.059	0.017
32.	Y, d, inf	0.029	**0.042**	0.027	0.029	**0.043**	0.027								
		Full	Rec	Exp											
Historical mean		0.025	0.073	0.017											

One-year stock returns in excess of the short-term interest rate $Y_t^{(R)}$ modelled with the nonparametric smoother and its linear counterpart (the single benchmarking approach) for several combinations of predictive variables X. In-sample period: 1872–1962, out-of-sample period: 1963–2022. Smallest five results (column-wise) highlighted in bold

Table 5 Comparison of predictive power: out-of-sample mean squared error for the full sample period (full), during recessions (rec) and expansions (exp)

Forecast combination			Non-par			Linear		
Nr.	Dim.	type	Full	Rec	Exp	Full	Rec	Exp
64.	All	Lasso	0.048	0.066	0.045	0.021	0.061	**0.015**
65.	All	Ridge	0.051	0.045	0.052	0.023	0.048	0.019
66.	All	Enet	0.056	0.052	0.056	0.021	0.051	0.016
67.	All	eLasso	0.031	0.057	0.027	0.021	0.048	0.017
68.	All	eRidge	0.025	0.058	0.019	0.023	0.050	0.019
69.	All	eEnet	0.025	0.057	0.020	0.021	0.052	0.017
70.	All	cENet	0.022	0.050	0.018	0.022	0.066	**0.015**
71.	All	Average	0.022	0.050	0.018	0.023	0.050	0.019
72.	1D	Lasso	**0.020**	0.056	0.014	0.022	0.067	**0.015**
73.	1D	Ridge	0.026	0.045	0.023	0.023	0.049	0.019
74.	1D	Enet	**0.020**	0.056	0.014	0.022	0.063	**0.015**
75.	1D	eLasso	**0.020**	0.059	0.014	0.023	0.057	0.017
76.	1D	eRidge	0.023	0.061	0.018	0.023	0.058	0.018
77.	1D	eEnet	**0.020**	0.059	0.014	0.023	0.057	0.017
78.	1D	cENet	0.023	0.060	0.017	0.023	0.058	0.018
79.	1D	Average	0.023	0.061	0.018	0.023	0.058	0.018
80.	2D	Lasso	0.022	0.064	0.015	0.021	0.062	**0.015**
81.	2D	Ridge	0.022	0.048	0.018	0.023	0.049	0.019
82.	2D	Enet	0.021	0.051	0.017	0.021	0.051	0.016
83.	2D	eLasso	0.022	0.054	0.017	0.021	0.056	**0.016**
84.	2D	eRidge	0.022	0.051	0.017	0.023	0.051	0.019
85.	2D	eEnet	0.023	0.057	0.018	0.021	0.054	0.016
86.	2D	cENet	0.022	0.051	0.017	0.022	0.066	**0.015**
87.	2D	Average	0.022	0.051	0.017	0.023	0.051	0.019
88.	3D	Lasso	0.051	0.067	0.048	0.021	0.056	**0.016**
89.	3D	Ridge	0.060	0.056	0.060	0.023	0.047	0.019
90.	3D	Enet	0.050	0.067	0.048	0.021	0.052	0.016
91.	3D	eLasso	0.049	0.064	0.047	**0.021**	0.049	0.016
92.	3D	eRidge	0.026	0.050	0.023	0.023	0.048	0.019
93.	3D	eEnet	0.049	0.064	0.047	0.021	0.052	0.016
94.	3D	cENet	0.022	0.058	0.016	0.031	0.046	0.029
95.	3D	Average	0.022	0.048	0.019	0.023	0.048	0.019

One-year stock returns in excess of the short-term interest rate $Y_t^{(R)}$ modelled by forecast combinations based on individual forecasts of models from Table 4 (the single benchmarking approach) applying the Lasso, the Ridge, the Enet, the eLasso, the eRidge, the eEnet, the cENet, and the simple average to all possible models or only to the k-dimensional ones with $k \in \{1, 2, 3\}$. In-sample period: 1872–1962, out-of-sample period: 1963–2022. Improvements (column-wise) compared to smallest measure in Table 4 highlighted in bold

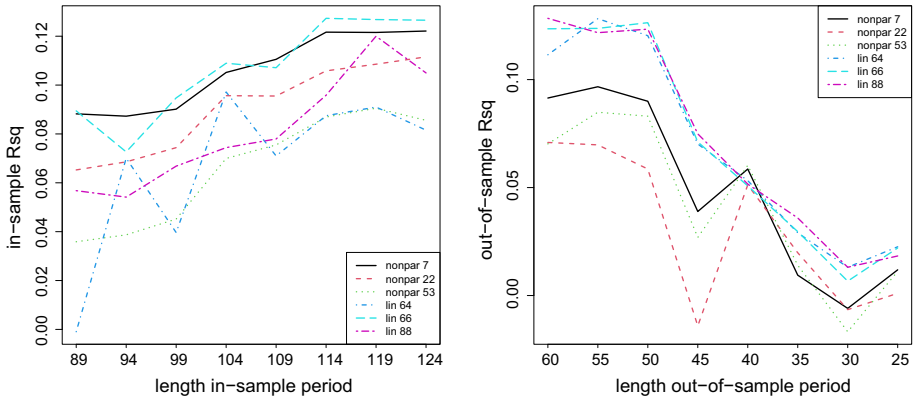

Fig. 3 Robustness over time (increasing in-sample period) for selected models for stock returns in excess of the risk-free rate. Left: R^2_V, Right: R^2_{oos}. Period: 1872–2022. Data: annual S&P 500

(instead of the 62 when single-benchmarking with the risk-free rate in Sect. 4.1) because all combinations which include inflation as a covariate are redundant. Table 6 presents the comparison of in-sample performance (measured by the R^2_V) and out-of-sample predictive power (measured by the R_{oos}) for the individual models. Based on the in-sample measure, the best five nonparametric models ($\{e, sp\}$, $\{Y, e, sp\}$, $\{r, sp\}$, $\{l, sp\}$, $\{r, l\}$) have a R^2_V in the range of 18.3–16.6%. We find again the term-spread to be included in most of these models. However, only the model $\{e, sp\}$ performs convincingly out of sample ($R_{oos} = 13.9\%$) as it is one of the five best predictive models ($\{d, e, sp\}$, $\{e, sp\}$, $\{e, r, sp\}$, $\{e, l, sp\}$, $\{e, r, l\}$) demonstrating a R^2_{oos} in the range of 13.9–13.0%. Note that the combination of real earnings and spread is included in four of those models. For the linear case, we find the same set of five best performing (in-sample) models with a R^2_V in the range of 18.5–16.9%. However, only two of those models beat the historical mean out-of-sample ($\{e, sp\}$, $\{Y, e, sp\}$ with an R^2_{oos} of 13.3% and 11.1%, resp.). The best five predictive models ($\{d, e, sp\}$, $\{e, sp\}$, $\{e, r, l\}$, $\{e, r, sp\}$, $\{e, l, sp\}$) include all the variable combination of real earnings and the spread, in most cases together with an additional covariate. Their R^2_{oos} values are in the range 13.4–12.9%.

When considering the correlation matrix of the in-sample forecasts from all one- and two-dimensional nonparametric models which is displayed in Fig. 4 (left hand side), one can observe that most predictors are highly positively correlated. The three exceptions are the models Y, sp, and $\{Y, sp\}$ which indeed are the models with the lowest R^2_V values, that is, the worst in-sample performance. The correlations for the out-of-sample forecasts shown in Fig. 4 (right hand side) are less pronounced but remain mostly positive (in contrast to the risk-free rate benchmark discussed above).

In the next step, we focus on the in- and out-of-sample performance of the 32 different forecast combination models. The corresponding results are shown in Table 7. When using all 40 available models, we find similar as before that the ENet (27.1%), the Ridge (23.8%), and the Lasso (23.7%) largely improve in-sample over the individual models. However, none of these forecast combinations produces forecasts with improved predictive power out-of-sample compared to individual models. We confirm our finding from the risk-free benchmark that those methods are prone to over-fitting. The restriction to one-dimensional models reduces a bit the in-sample R^2_V (ENet: 21.8%, Ridge: 20.1%, Lasso: 19.2%) but all of these models have now R^2_{oos} values larger than 12.7%. However, none of the forecast

Table 6 Comparison of predictive power: in-sample (measured by the R^2_V) versus out-of-sample (measured by the R^2_{oos})

Model		Non-par		Linear		Model		Non-par		Linear	
Nr.	X	R^2_V	R^2_{oos}	R^2_V	R^2_{oos}	Nr.	X	R^2_V	R^2_{oos}	R^2_V	R^2_{oos}
1.	Y	− 0.023	0.008	− 0.022	0.007	33.	Y, d, sp	0.154	0.121	0.154	0.093
2.	d	0.139	− 0.028	0.141	− 0.027	34.	Y, e, r	0.161	− 0.196	0.165	− 0.200
3.	e	0.159	− 0.025	0.160	− 0.024	35.	Y, e, l	0.143	− 0.173	0.148	− 0.186
4.	r	0.102	− 0.009	0.103	− 0.009	36.	Y, e, inf	–	–	–	–
5.	l	0.131	0.034	0.133	0.034	37.	Y, e, sp	**0.174**	0.120	**0.178**	0.111
6.	inf	–	–	–	-	38.	Y, r, l	0.145	− 0.032	0.148	− 0.028
7.	sp	0.055	0.074	0.038	− 0.060	39.	Y, r, inf	–	–	–	–
8.	Y, d	0.135	− 0.065	0.137	− 0.068	40.	Y, r, sp	0.146	0.032	0.148	− 0.028
9.	Y, e	0.158	− 0.066	0.160	− 0.069	41.	Y, l, inf	–	–	–	–
10.	Y, r	0.087	− 0.018	0.088	− 0.021	42.	Y, l, sp	0.146	0.037	0.148	− 0.028
11.	Y, l	0.117	0.024	0.118	0.022	43.	Y, inf, sp	–	–	–	–
12.	Y, inf	–	–	–	-	44.	d, e, r	0.123	−0.230	0.127	− 0.200
13.	Y, sp	0.030	− 0.070	0.011	− 0.058	45.	d, e, l	0.110	− 0.101	0.116	− 0.070
14.	d, e	0.129	− 0.026	0.133	− 0.024	46.	d, e, inf	–	–	–	–
15.	d, r	0.137	− 0.275	0.140	− 0.249	47.	d, e, sp	0.152	**0.139**	0.158	**0.134**
16.	d, l	0.123	− 0.186	0.125	− 0.146	48.	d, r, l	0.147	0.085	0.153	0.060
17.	d, inf	–	–	–	-	49.	d, r, inf	–	–	–	–
18.	d, sp	0.164	0.122	0.165	0.113	50.	d, r, sp	0.149	0.101	0.153	0.060
19.	e, r	0.153	−0.097	0.157	− 0.092	51.	d, l, inf	–	–	–	–

Table 6 continued

Model		Non-par		Linear		Model		Non-par		Linear	
Nr.	X	R^2_V	R^2_{oos}	R^2_V	R^2_{oos}	Nr.	X	R^2_V	R^2_{oos}	R^2_V	R^2_{oos}
20.	e, l	0.142	− 0.059	0.146	− 0.054	52.	d, l, sp	0.149	0.100	0.153	0.060
21.	e, inf	–	–	–	-	53.	d, inf, sp	-	-	-	-
22.	e, sp	**0.183**	**0.139**	**0.185**	**0.133**	54.	e, r, l	0.161	**0.130**	0.168	**0.129**
23.	r, l	**0.166**	− 0.021	**0.169**	−0.022	55.	e, r, inf	–	–	–	–
24.	r, inf	–	–	–	-	56.	e, r, sp	0.163	**0.137**	0.168	**0.129**
25.	r, sp	**0.168**	0.006	**0.169**	-0.022	57.	e, l, inf	-	-	-	-
26.	l, inf	–	–	–	-	58.	e, l, sp	0.163	**0.137**	0.168	**0.129**
27.	l, sp	**0.168**	0.007	**0.169**	-0.022	59.	e, inf, sp	–	–	–	–
28.	inf, sp	–	–	–	-	60.	r, l, inf	–	–	–	–
29.	Y, d, e	0.130	− 0.064	0.134	−0.069	61.	r, l, sp	–	–	–	–
30.	Y, d, r	0.149	− 0.537	0.152	−0.503	62.	r, inf, sp	–	–	–	–
31.	Y, d, l	0.133	− 0.844	0.135	−0.789	63.	l, inf, sp	–	–	–	–
32.	Y, d, inf	–	–	–	-						

One-year stock returns in excess of the inflation rate $Y_t^{(C)}$ modelled with the nonparametric smoother and its linear counterpart (the double benchmarking approach) for several combinations of predictive variables X. In-sample period: 1872–1962, out-of-sample period: 1963–2022. Largest five results (column-wise) highlighted in bold

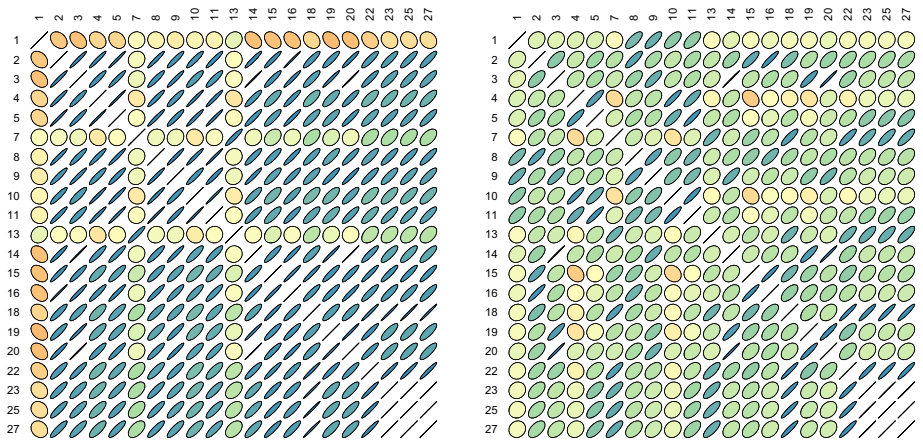

Fig. 4 Correlations of predictions for stock returns in excess of the inflation rate (for nonlinear models of one or two predictive variables). Left: in-sample, Right: out-of-sample. period: 1872–2022. Data: annual S&P 500

combination models is able to improve out-of-sample over the best individual model. Note that the cENet seems to perform reasonable under the double inflation benchmark. The reason is that it selects only the single two-dimensional model based on real earnings and the term spread, one of the best individual candidate models. For the linear counterpart, the situation is similar. In-sample, mostly the ENet improves over the individual models. But out-of-sample, non of the 32 ways of combining individual forecasts was able to increase the predictive power over the best individual three-dimensional model.

For the performance during recessions or economic expansions, we find similar pattern as under the risk-free rate benchmark. Table 9 shows the mean squared errors for the individual models and Table 8 for the forecast combination models. One observes again that the (nonparametric and linear) models with the smallest out-of-sample mean squared error over the full period excellently perform during economic expansions but are only slightly able to beat the historical mean during the recessions. There are again several models which are performing very promising during recessions (for example, $\{Y, d, l\}$ or $\{Y, d, r\}$). During the economic expansions, however, these models cannot beat the historical mean at all.

We finish the empirical part with the robustness check over time. Figure 5 shows the R^2_V (left) and the R^2_{oos} (right) for models with the largest out-of-sample R^2 (we restrict ourselves again to the best three nonparametric and the best three linear models, resp.). In contrast to the risk-free benchmark, the performance over time is more stable under the inflation benchmark. For the nonparametric models, 8 of the individual models and 18 of the forecast combinations were able to beat the historical mean in each setting. However, the three best performing models are $\{e, sp\}$, $\{d, e, sp\}$, and $\{e, l, sp\}$. For the linear models, a similar set of 8 individual models and 28 of the forecast combinations yield positive R^2_{oos} values over time. Again, the three best performing models were individual models: $\{d, e, sp\}$, $\{e, r, l\}$, and $\{e, l, sp\}$. Figure 5 also shows that the in-sample performance is now quite stable over time. However, the out-of-sample measure sharply drops at the end of the considered period. Note also that nonparametric and linear models are close together in-sample as well as out-of-sample. Nevertheless, the nonparametric models based on the covariates $\{e, sp\}$ and $\{d, e, sp\}$ performed best in terms of largest and stable R^2_{oos}, and are thus the preferred model choices.

Table 7 Comparison of predictive power: in-sample (measured by the R^2_V) versus out-of-sample (measured by the R^2_{oos})

Forecast combination			Non-par		Linear	
Nr.	Dim.	type	R^2_V	R^2_{oos}	R^2_V	R^2_{oos}
64.	All	Lasso	**0.237**	0.119	0.164	0.115
65.	All	Ridge	**0.238**	0.089	0.180	0.113
66.	All	Enet	**0.271**	0.097	**0.199**	0.113
67.	All	eLasso	0.171	0.092	0.178	0.111
68.	All	eRidge	0.165	0.029	0.171	0.103
69.	All	eEnet	0.169	0.100	0.175	0.120
70.	All	cENet	0.154	0.063	0.177	0.055
71.	All	Average	0.171	0.099	0.171	0.103
72.	1D	Lasso	**0.192**	0.129	0.182	0.113
73.	1D	Ridge	**0.201**	0.129	0.171	0.110
74.	1D	Enet	**0.218**	0.127	**0.209**	0.111
75.	1D	eLasso	0.119	0.079	0.113	0.082
76.	1D	eRidge	0.141	0.057	0.138	0.064
77.	1D	eEnet	0.119	0.079	0.113	0.082
78.	1D	cENet	0.159	−0.025	0.156	0.107
79.	1D	Average	0.141	0.057	0.138	0.064
80.	2D	Lasso	**0.236**	0.037	0.153	0.116
81.	2D	Ridge	**0.228**	0.091	0.184	0.115
82.	2D	Enet	**0.273**	0.036	**0.212**	0.109
83.	2D	eLasso	0.156	0.021	**0.188**	0.085
84.	2D	eRidge	0.171	0.092	0.170	0.098
85.	2D	eEnet	0.159	0.030	0.166	0.091
86.	2D	cENet	0.143	−0.055	0.180	0.082
87.	2D	Average	0.171	0.092	0.170	0.098
88.	3D	Lasso	**0.263**	0.127	**0.196**	0.114
89.	3D	Ridge	**0.260**	0.124	**0.210**	0.108
90.	3D	Enet	**0.274**	0.123	**0.212**	0.107
91.	3D	eLasso	0.154	0.121	0.180	0.027
92.	3D	eRidge	0.168	0.124	0.171	0.107
93.	3D	eEnet	0.168	0.124	0.178	0.089
94.	3D	cENet	0.022	0.000	0.155	− 0.015
95.	3D	Average	0.168	0.105	0.171	0.107

One-year stock returns in excess of the inflation rate $Y_t^{(C)}$ modelled by forecast combinations based on individual forecasts of models from Table 6 (the double benchmarking approach) applying the Lasso, the Ridge, the Enet, the eLasso, the eRidge, the eEnet, the cENet, and the simple average to all possible models or only to the k-dimensional ones with $k \in \{1, 2, 3\}$. In-sample period: 1872–1962, out-of-sample period: 1963–2022. Improvements (column-wise) compared to largest measure in Table 2 highlighted in bold

Table 8 Comparison of predictive power: out-of-sample mean squared error for the full sample period (full), during recessions (rec) and expansions (exp)

Forecast combination			Non-par			Linear		
Nr.	Dim.	type	Full	Rec	Exp	Full	Rec	Exp
64.	All	Lasso	0.021	0.052	0.016	0.021	0.051	0.016
65.	All	Ridge	0.021	0.051	0.017	0.021	0.051	0.016
66.	All	Enet	0.021	0.052	0.016	0.021	0.048	0.017
67.	All	eLasso	0.021	0.050	0.017	0.021	0.046	0.017
68.	All	eRidge	0.023	0.049	0.019	0.021	0.052	0.016
69.	All	eEnet	0.021	0.053	0.016	0.021	0.052	0.016
70.	All	cENet	0.022	0.047	0.018	0.022	0.047	0.018
71.	All	Average	0.021	0.052	0.016	0.021	0.052	0.016
72.	1D	Lasso	0.020	0.048	0.016	0.021	0.046	0.01
73.	1D	Ridge	0.020	0.054	0.015	0.021	0.055	0.016
74.	1D	Enet	0.021	0.046	0.017	0.021	0.046	0.017
75.	1D	eLasso	0.022	0.068	0.015	0.022	0.069	**0.014**
76.	1D	eRidge	0.022	0.066	0.015	0.022	0.067	0.015
77.	1D	eEnet	0.022	0.068	0.015	0.022	0.069	**0.014**
78.	1D	cENet	0.024	0.051	0.020	0.021	0.063	0.014
79.	1D	Average	0.022	0.066	0.015	0.022	0.067	0.015
80.	2D	Lasso	0.023	0.052	0.018	0.021	0.055	0.016
81.	2D	Ridge	0.021	0.053	0.017	0.021	0.054	0.016
82.	2D	Enet	0.023	0.052	0.018	0.021	0.044	0.017
83.	2D	eLasso	0.023	0.053	0.018	0.022	0.047	0.018
84.	2D	eRidge	0.021	0.054	0.016	0.021	0.054	0.016
85.	2D	eEnet	0.023	0.054	0.018	0.021	0.053	0.016
86.	2D	cENet	0.025	0.046	0.022	0.022	0.046	0.018
87.	2D	Average	0.021	0.054	0.016	0.021	0.054	0.016
88.	3D	Lasso	0.021	0.050	0.016	0.021	0.048	0.017
89.	3D	Ridge	0.021	0.053	0.016	0.021	0.050	0.017
90.	3D	Enet	0.021	0.054	0.015	0.021	0.044	0.017
91.	3D	eLasso	0.021	0.049	0.016	0.023	0.039	0.020
92.	3D	eRidge	0.021	0.052	0.016	0.021	0.047	0.017
93.	3D	eEnet	0.021	0.052	0.016	0.021	0.043	0.018
94.	3D	cENet	0.021	0.047	0.017	0.024	0.041	0.021
95.	3D	Average	0.021	0.047	0.017	0.021	0.047	0.017

One-year stock returns in excess of the inflation rate $Y_t^{(C)}$ modelled by forecast combinations based on individual forecasts of models from Table 4 (the single benchmarking approach) applying the Lasso, the Ridge, the Enet, the eLasso, the eRidge, the eEnet, the cENet, and the simple average to all possible models or only to the k-dimensional ones with $k \in \{1, 2, 3\}$. In-sample period: 1872–1962, out-of-sample period: 1963–2022. Improvements (column-wise) compared to smallest measure in Table 4 highlighted in bold

Table 9 Comparison of predictive power: out-of-sample mean squared error for the full sample period (full), during recessions (rec) and expansions (exp)

Model		Non-par			Linear			Model		Non-par			Linear		
Nr.	*X*	Full	Rec	Exp	Full	Rec	Exp	Nr.	*X*	Full	Rec	Exp	Full	Rec	Exp
1.	Y	0.023	0.081	**0.014**	0.023	0.081	**0.014**	33.	Y, d, sp	0.021	0.049	0.016	0.021	0.048	0.017
2.	d	0.024	0.051	0.020	0.024	0.051	0.020	34.	Y, e, r	0.028	**0.034**	0.027	0.028	**0.033**	0.027
3.	e	0.024	0.051	0.020	0.024	0.051	0.020	35.	Y, e, l	0.028	0.039	0.026	0.028	**0.038**	0.026
4.	r	0.024	0.074	0.016	0.024	0.074	0.016	36.	Y, e, inf	–	–	–	–	–	–
5.	l	0.023	0.071	**0.015**	0.023	0.071	**0.015**	37.	Y, e, sp	0.021	0.046	0.017	0.021	0.046	0.017
6.	inf	–	–	–	–	–	–	38.	Y, r, l	0.024	0.066	0.018	0.024	0.066	0.018
7.	sp	0.022	0.076	**0.013**	0.025	0.078	0.017	39.	Y, r, inf	–	–	–	–	–	–
8.	Y, d	0.025	0.047	0.022	0.025	0.046	0.022	40.	Y, r, sp	0.023	0.065	0.016	0.024	0.066	0.018
9.	Y, e	0.025	0.045	0.022	0.025	0.045	0.022	41.	Y, l, inf	–	–	–	–	–	–
10.	Y, r	0.024	0.071	0.017	0.024	0.071	0.017	42.	Y, l, sp	0.023	0.065	0.016	0.024	0.066	0.018
11.	Y, l	0.023	0.069	0.016	0.023	0.069	0.016	43.	Y, inf, sp	–	–	–	–	–	–
12.	Y, inf	–	–	–	–	–	–	44.	d, e, r	0.029	**0.036**	0.028	0.028	**0.037**	0.027
13.	Y, sp	0.025	0.076	0.017	0.025	0.077	0.017	45.	d, e, l	0.026	0.044	0.023	0.025	0.047	0.022
14.	d, e	0.024	0.050	0.020	0.024	0.051	0.020	46.	d, e, inf	–	–	–	–	–	–
15.	d, r	0.030	**0.038**	0.029	0.029	0.038	0.028	47.	d, e, sp	**0.020**	0.050	0.016	**0.020**	0.051	0.016
16.	d, l	0.028	0.042	0.026	0.027	0.044	0.024	48.	d, r, l	0.022	0.058	0.016	0.022	0.061	0.016
17.	d, inf	–	–	–	–	–	–	49.	d, r, inf	–	–	–	–	–	–
18.	d, sp	0.021	0.052	0.016	0.021	0.052	0.016	50.	d, r, sp	0.021	0.058	0.015	0.022	0.061	0.016
19.	e, r	0.026	0.042	0.023	0.026	0.042	0.023	51.	d, l, inf	–	–	–	–	–	–
20.	e, l	0.025	0.048	0.021	0.025	0.048	0.021	52.	d, l, sp	0.021	0.058	0.015	0.022	0.061	0.016
21.	e, inf	–	–	–	–	–	–	53.	d, inf, sp	–	–	–	–	–	–

Table 9 continued

Model		Non-par			Linear			Model		Non-par			Linear		
Nr.	X	Full	Rec	Exp	Full	Rec	Exp	Nr.	X	Full	Rec	Exp	Full	Rec	Exp
22.	e, sp	**0.020**	0.051	0.016	**0.020**	0.051	0.016	54.	e, r, l	**0.020**	0.052	0.016	**0.020**	0.052	**0.016**
23.	r, l	0.024	0.068	0.017	0.024	0.068	0.017	55.	e, r, inf	–	–	–	–	–	–
24.	r, inf	–	–	–	–	–	–	56.	e, r, sp	**0.020**	0.052	**0.015**	**0.020**	0.052	**0.016**
25.	r, sp	0.023	0.068	0.017	0.024	0.068	0.017	57.	e, l, inf	–	–	–	–	–	–
26.	l, inf	–	–	–	–	–	–	58.	e, l, sp	**0.020**	0.052	**0.015**	**0.020**	0.052	**0.016**
27.	l, sp	0.023	0.068	0.017	0.024	0.068	0.017	59.	e, inf, sp	–	–	–	–	–	–
28.	inf, sp	–	–	–	–	–	–	60.	r, l, inf	–	–	–	–	–	–
29.	Y, d, e	0.025	0.046	0.022	0.025	0.045	0.022	61.	r, l, sp	–	–	–	–	–	–
30.	Y, d, r	0.036	**0.032**	0.037	0.035	**0.031**	0.036	62.	r, inf, sp	–	–	–	–	–	–
31.	Y, d, l	0.043	**0.029**	0.046	0.042	**0.029**	0.044	63.	l, inf, sp	–	–	–	–	–	–
32.	Y, d, inf	–	–	–	–	–	–								
		Full	Rec	Exp											
Historical mean		0.024	0.080	0.015											

One-year stock returns in excess of the inflation rate $Y_t^{(C)}$ modelled with the nonparametric smoother and its linear counterpart (the double benchmarking approach) for several combinations of predictive variables X. In-sample period: 1872–1962, out-of-sample period: 1963–2022. Smallest five results (column-wise) highlighted in bold

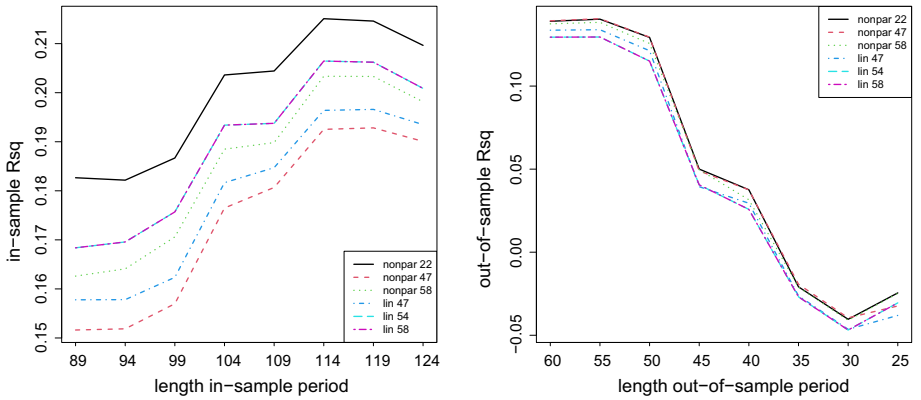

Fig. 5 Robustness over time (increasing in-sample period) for selected models for stock returns in excess of the inflation rate. Left: R^2_V, Right: R^2_{oos}. Period: 1872–2022. Data: annual S&P 500

5 Conclusion

In this paper, we analyse whether forecast combinations of stock returns in excess of different benchmarks are able to improve over the individual models. Our focus lies thereby on nonlinear predictive functions estimated by a fully nonparametric smoother with the covariates and smoothing parameters chosen by cross-validation. We extent the approach of Kyriakou et al. (2021a) to three-dimensional models and find for some of them a reasonable performance both in-sample and out-of-sample. However, the low number of observations in the estimation sample limits the complexity of the fitted models. This reduces the probability of choosing such models as the in-sample measure is worse when compared to simpler models of lower dimensionality. Note that the reason lies in the fact that the rate of convergence of the local-linear smoother is slower than, for example, in parametric regression (Hansen, 2008a).

We find further that the classical shrinkage methods (Lasso, Ridge, ENet) are prone to in-sample over-fitting when all individual forecasts are used as possible candidates. As a consequence, the suggested predictive power is spurious and the out-of-sample performance is indeed very poor. The restriction to one-dimensional candidates helps to balance in-sample and out-of-sample behaviour and improves the out-of-sample predictive power. We also find that forecast combinations perform better than the simple historical average. However, the individual nonparametric models outperform linear models and combination forecasts throughout. Finally, the double inflation benchmark results in a more stable performance compared to the single risk-free benchmark.

Recently, there has been a fast growth of methodology to process data for financial applications. This again provides us with the challenge of making sure that more and more complex methodology is indeed also better than simpler methods. It is well-known that complexity often comes with a price. With this current study, we get to the conclusion that a simple benchmark methodology is indeed as good as a selected collection of the most popular, but also more complex and less transparent, modern ML-type approaches. Also for the financial practitioner, implementing the model and communicating results are simply easier carried out with a simpler methodology. So, it is important for policy making and financial planning of long-term saving that complexity and lack of transparency are only introduced to the econometric modelling when it is absolutely necessary. In the important challenge of

understanding long-term financial returns based on econometric modelling, the conclusion of this study seems to be that complexity does not pay off well enough and that it is better to use simpler benchmark methods.

Funding Open access funding provided by University of Klagenfurt.

References

Aiolfi, M., & Timmermann, M. (2006). Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics, 135,* 31–53. https://doi.org/10.1016/j.jeconom.2005.07.015.

Akyildirim, E., Bariviera, A.F., Nguyen, D.K., & Sensoy, A. (2022). Forecasting high-frequency stock returns: A comparison of alternative methods. *Annals of Operations Research*. https://doi.org/10.1007/s10479-021-04464-8.

Akyildirim, E., Goncu, A., & Sensoy, A. (2021). Prediction of cryptocurrency returns using machine learning. *Annals of Operations Research, 297,* 3–36. https://doi.org/10.1007/s10479-020-03575-y.

Asimit, V., Kyriakou, I., & Nielsen, J.P. (2020). Special Issue "Machine Learning in Insurance". *Risks*. https://doi.org/10.3390/risks8020054.

Athey, S., & Imbens, G. (2019). Machine learning methods economists should know about. *arXiv*

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Journal of the Operational Research Society, 20,* 451–468. https://doi.org/10.1057/jors.1969.103.

Caldeira, J. F., Gupta, R., Torrent, H., & Forecasting, U. S. (2020). Aggregate stock market excess return: Do functional data analysis add economic value? *Mathematics, 8,* 1–16. https://doi.org/10.3390/math8112042.

Campbell, J., & Shiller, R. (1988). The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies, 1,* 195–228.

Campbell, J., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies, 21,* 1509–1531. https://doi.org/10.1093/rfs/hhm055.

Chen, Q., & Hong, Y. (2010). *Predictability of equity returns over different time horizons: a nonparametric approach*. Cornell University/Department of Economics. Working Paper.

Cheng, T., Gao, J., & Linton, O. (2019). *Nonparametric predictive regressions for stock return predictions*. Cambridge Working Papers in Economics: 1932

Claeskens, G., Magnus, J., Vasnev, A., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting, 32,* 754–762. https://doi.org/10.1016/j.ijforecast.2015.12.005.

Coqueret, G., & Guida, T. (2020). Training trees on tails with applications to portfolio choice. *Annals of Operations Research, 288,* 181–221. https://doi.org/10.1007/s10479-020-03539-2.

Coulombe, P., Leroux, M., Stevanovic, D., & Surprenant, S. (2020). How is machine learning useful for macroeconomic forecasting? *arXiv*

Diebold, F., & Shin, M. (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives. *International Journal of Forecasting, 35,* 1679–1691. https://doi.org/10.1016/j.ijforecast.2018.09.006.

Dixon, M., Halperin, I., & Bilokon, P. (2020). *Machine Learning in Finance*. Cham: Springer. https://doi.org/10.1007/978-3-030-41068-1.

Elliott, G., Gargano, A., & Timmermann, A. (2013). Complete subset regressions. *Journal of Econometrics, 177,* 357–373. https://doi.org/10.1016/j.jeconom.2013.04.017.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance, 25,* 383–417. https://doi.org/10.2307/2325486.

Fama, E. F. (1988). Dividend yields and expected stock returns. *Journal of Financial Economics, 22,* 3–25.

Freyberger, J., Neuhierl, A., & Weber, M. (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies, 5,* 2326–2377. https://doi.org/10.1093/rfs/hhz123.

Gambella, C., Ghaddar, B., & Naoum-Sawaya, J. (2021). Optimization problems for machine learning: A survey. *European Journal of Operational Research, 290,* 807–828. https://doi.org/10.1016/j.ejor.2020.08.045.

Gerrard, R., Hiabu, M., Kyriakou, I., & Nielsen, J. P. (2018). Self-selection and risk sharing in a modern world of life-long annuities. *British Actuarial Journal, 23,* e30. https://doi.org/10.1017/S135732171800020X.

Gerrard, R., Hiabu, M., Kyriakou, I., & Nielsen, J. P. (2019). Communication and personal selection of pension saver's financial risk. *European Journal of Operational Research, 274,* 1102–1111.

Gerrard, R., Hiabu, M., Nielsen, J. P., & Vodička, P. (2020). Long-term real dynamic investment planning. *Insurance: Mathematics and Economics, 92,* 90–103. https://doi.org/10.1016/j.insmatheco.2020.03.002.

Granger, C. W. J., & Jeon, Y. (2004). Thick modeling. *Economic Modelling, 2,* 323–343. https://doi.org/10.1016/S0264-9993(03)00017-8.

Granger, C. W. J., & Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting, 3,* 197–204. https://doi.org/10.1002/for.3980030207.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies, 33,* 2223–2273. https://doi.org/10.1093/rfs/hhaa009.

Guerrero, V., & Pena, D. (2003). Combining multiple time series predictors: A useful inferential procedure. *Journal of Statistical Planning and Inference, 1,* 249–276. https://doi.org/10.1016/S0378-3758(02)00186-6.

Hansen, B. (2008a). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory, 24,* 726–748.

Hansen, B. (2008b). Least-squares forecast averaging. *Journal of Econometrics, 2,* 342–350. https://doi.org/10.1016/j.jeconom.2008.08.022.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning*. Cham: Springer.

Hiabu, M., Mammen, E., & Meyer, J. (2020). Random planted forest: A directly interpretable tree ensemble. *arXiv*

Hoerl, A., & Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 1,* 55–67. https://doi.org/10.1080/00401706.1970.10488634.

Iworiso, J., & Vrontos, S. (2020). On the directional predictability of equity premium using machine learning techniques. *Journal of Forecasting, 39,* 449–469. https://doi.org/10.1002/for.2632.

Kyriakou, I., Mousavi, P., Nielsen, J. P., & Scholz, M. (2020). Longer-term forecasting of excess stock returns-the five-year case. *Mathematics, 8,* 1–20.

Kyriakou, I., Mousavi, P., Nielsen, J. P., & Scholz, M. (2021a). Forecasting benchmarks of long-term stock returns via machine learning. *Annals of Operations Research, 287,* 221–240. https://doi.org/10.1007/s10479-019-03338-4.

Kyriakou, I., Mousavi, P., Nielsen, J. P., & Scholz, M. (2021b). Short-term exuberance and long-term stability: A simultaneous optimization of stock return predictions for short and long horizons. *Mathematics, 9,* 1–19.

Lee, Y., & Sul, D. (2021). Depth-weighted forecast combination: Application to COVID-19 cases. *CPR Working Papers, Paper No. 238*

Lettau, M., & Van Nieuwerburgh, S. (2008). Reconciling the return predictability evidence. *Review of Financial Studies, 21,* 1601–1652.

Lin, H., Wu, C., & Zhou, G. (2018). Forecasting corporate bond returns with a large set of predictors: An iterated combination approach. *Management Science, 9,* 4218–4238. https://doi.org/10.1287/mnsc.2017.2734.

Lioui, A., & Poncet, P. (2019). Long horizon predictability: An asset allocation perspective. *European Journal of Operational Research, 278,* 961–975. https://doi.org/10.1016/j.ejor.2019.04.040.

Mammen, E., Nielsen, J. P., Scholz, M., & Sperlich, S. (2019). Conditional variance forecasts for long-term stock returns. *Risks, 7,* 1–22.

Mascio, D., Fabozzi, F., & Zumwalt, J. K. (2020). Market timing using combined forecasts and machine learnings. *Journal of Forecasting, 40,* 1–16. https://doi.org/10.1002/for.2690.

Merton, R. C. (2014). The crisis in retirement planning. *Harvard Business Review, 92,* 43–50.

Nielsen, J. P., & Sperlich, S. (2003). Prediction of stock returns: A new way to look at it. *ASTIN Bulletin, 2,* 399–417.

Oztekin, A., Kizilaslan, R., Freund, S., & Iseri, A. (2016). A data analytic approach to forecasting daily stock returns in an emerging market. *European Journal of Operational Research, 3,* 697–710. https://doi.org/10.1016/j.ejor.2016.02.056.

Qian, W., Rolling, C., Cheng, G., & Yang, Y. (2019). On the forecast combination puzzle. *Econometrics*. https://doi.org/10.3390/econometrics7030039.

Rapach, D., & Zhou, G. (2020). *Time-series and cross-sectional stock return forecasting: new machine learning methods* (pp. 1–34). Machine Learning for Asset Management: New Developments and Financial Applications.

Rapach, D., Strauss, J., & Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies, 23,* 821–862. https://doi.org/10.1093/rfs/hhp063.

Scholz, M., Nielsen, J., & Sperlich, S. (2015). Nonparametric prediction of stock returns based on yearly data: The long-term view. *Insurance: Mathematics and Economics, 65,* 143–155.

Scholz, M., Sperlich, S., & Nielsen, J. (2016). Nonparametric long term prediction of stock returns with generated bond yields. *Insurance: Mathematics and Economics, 69,* 82–96.

Shiller, R. (1989). *Market volatility*. Cambridge, MA: MIT Press.

Stambaugh, R. (1999). Predictive regressions. *Journal of Financial Economics, 54,* 375–421.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological), 1,* 267–288.

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. Granger, & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (Vol. 1, pp. 135–196).

Wu, W., Chen, J., Xu, L., He, Q., & Tindall, M. (2020). A statistical learning approach for stock selection in the Chinese stock market. *Financial Innovation, 5,* 1–18. https://doi.org/10.1186/s40854-019-0137-1.

Yang, J., Cabrera, J., & Wang, T. (2010). Nonlinearity, data-snooping, and stock index ETF return predictability. *European Journal of Operational Research, 200,* 498–507. https://doi.org/10.1016/j.ejor.2009.01.009.

Zhang, H., He, Q., Jacobsen, B., & Jiang, F. (2020). Forecasting stock returns with model uncertainty and parameter instability. *Journal of Applied Econometric, 35,* 629–644. https://doi.org/10.1002/jae.2747.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101,* 1418–1429. https://doi.org/10.1198/016214506000000735.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67,* 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.