Yae and Luo *Financial Innovation* (2023) 9:94
https://doi.org/10.1186/s40854-023-00497-z

Financial Innovation

RESEARCH Open Access

Robust monitoring machine: a machine learning solution for out-of-sample R_2-hacking in return predictability monitoring

James Yae[1*] and Yang Luo[1]

*Correspondence:
syae.uh@gmail.com

[1] C. T. Bauer College of Business, University of Houston, 4750 Calhoun Rd, Houston, TX 77204, USA

Abstract

The out-of-sample R^2 is designed to measure forecasting performance without look-ahead bias. However, researchers can hack this performance metric even without multiple tests by constructing a prediction model using the intuition derived from empirical properties that appear only in the test sample. Using ensemble machine learning techniques, we create a virtual environment that prevents researchers from peeking into the intuition in advance when performing out-of-sample prediction simulations. We apply this approach to robust monitoring, exploiting a dynamic shrinkage effect by switching between a proposed forecast and a benchmark. Considering stock return forecasting as an example, we show that the resulting robust monitoring forecast improves the average performance of the proposed forecast by 15% (in terms of mean-squared-error) and reduces the variance of its relative performance by 46% while avoiding the out-of-sample R^2-hacking problem. Our approach, as a final touch, can further enhance the performance and stability of forecasts from any models and methods.

Keywords: Machine learning, Out-of-sample R^2-hacking, Return predictability, Monitoring

JEL Classification: C52, C53, C55, C58, G17

Introduction

The out-of-sample R^2 is no better than in-sample R^2 regarding data snooping concerns. Persistent researchers can attempt multiple prediction models and ingenuously report only the best performing one (Inoue and Kilian 2005, 2006, de Prado 2019). Alternatively, careful researchers examine a model of their own choice, but only the lucky ones obtain false positive results that are good enough for publication (Chordia et al. 2017). However, the problem, in reality, is even deeper than that. Researchers often construct a prediction model using the intuition derived from recent empirical findings that did not exist in the training sample period (Yae Forthcoming). It is unlikely that prior to the recent findings, forecasters chose such a model without a hint from the future; that is, an unintended look-ahead bias arises in pseudo out-of-sample testing.

For example, some early studies, such as Pesaran and Timmermann (1995), show that stock return predictability is time-varying: the predictability is stronger in recessions than in expansions. Then, many follow-up studies utilize such empirical facts to improve out-of-sample predictability further.[1] The truth is, however, a hypothetical forecaster in out-of-sample prediction simulations would not choose such stylized models without sufficient evidence at the moment of prediction (Martin and Nagel 2022). Therefore, the best option previously available for the forecaster is to consider all possible models and choose one or a combination, ex-ante optimally, without help of the not-yet-available intuition.

We consider a machine learning approach to tackle this comprehensive out-of-sample R^2-hacking problem. However, our strategy differs from others, which highlight the superior prediction accuracy of machine learning. We find a new solution by exploiting the weakness of machine learning: the black-box-like nature that potentially exacerbates the out-of-sample R^2-hacking problem and blurs economic intuition.[2] Figuratively, using this *black box*, we create a virtual environment that prevents researchers from peeking into the intuition from the future when they perform out-of-sample prediction simulations.[3] Therefore, we do not attempt to maximize the average forecasting performance ex-post. Instead, we aim to measure the attainable level of out-of-sample predictability in practice using machine learning.

As a practical example, we demonstrate our solution for out-of-sample R^2-hacking in the context of robust monitoring forecasts; that is, a forecast that switches between a proposed forecast and a benchmark while utilizing the conditional predictabilities, which are monitored in real-time.[4] To avoid out-of-sample R^2-hacking, we suggest the following three-step approach: (1) choose an information set (i.e., conditioning variables) for monitoring, independently of intuition available only after the moment of prediction, (2) approximate the entire time-series model space of the conditioning variables with feature engineering, and (3) apply a combination of ensemble machine learning algorithms, instead of choosing what works best ex-post after attempting many. That is, we design each step to exclude the out-of-sample R^2-hacking associated with selecting data, models, and optimization algorithms, respectively. Therefore, implementing the first two steps are as important as the last step, which ensembles multiple machine learning algorithms. We call this class of machine learning approach for robust monitoring forecasts, *robust monitoring machine*.

We apply our robust monitoring machine to a real-world prediction example for demonstration purposes. The proposed forecast in this example is a combination forecast from Rapach et al. (2010), which set an equal-weighted average of the 14 stock market predictors studied in Goyal and Welch (2008). The benchmark forecast is a real-time historical average of stock market returns, as Campbell and Thompson (2008) suggests,

[1] Among many others, Dangl and Halling (2012) study time-varying coefficient models whose coefficients follow a random walk. Henkel et al. (2011) and Zhu and Zhu (2013) choose regime switching models instead. Rapach et al. (2013) have the discounted mean squared forecast error (DMSFE) in the past to determine the combination weights, similarly to Stock and Watson (2004) and Bates and Granger (1969).

[2] See (Prado and López 2018) and (Bailey et al. 2015) for out-of-sample R^2-hacking problem.

[3] Similarly, Kou et al. (2022) propose an extension of group decision-making and spherical fuzzy numbers.

[4] The method of monitoring forecasts solves a classification problem by supervised learning. For recent development in clustering (unsupervised learning) with financial data, please refer to Kou et al. (2014) and Li et al. (2021).

while defining the out-of-sample R^2. Then, we apply the three-step approach as follows. First, we choose the smallest possible information set for monitoring: a history of the prediction loss (e.g., squared-error loss) differences between the proposed and benchmark forecast. Next, following Christ et al. (2018), we convert the variable in our information set into hundreds of time-series features, which act as building blocks to approximate the entire time-series model space. Finally, we blindly combine (with equal weights) three popular ensemble machine learning algorithms that we sequentially trained and validated for out-of-sample forecasting: random forest, extremely randomized trees, and gradient boosting. We stress that we rule out potential out-of-sample R^2-hacking problem by *intentionally* performing these tasks in an unsophisticated manner.

Nevertheless, the out-of-sample classification performance of the robust monitoring machine remains outstanding, despite our efforts to avoid out-of-sample R^2-hacking problems. Its sensitivity is 57.0% and specificity is 56.8% without any look-ahead biases, which statistically differ from those of uninformative random classifiers. Our results imply that the past return-predictability truly contains information for the future return predictability of proposed and benchmark forecasts.

To quantify the advantages of the robust monitoring machine, we measure a prediction loss difference (between a forecast and the benchmark forecast). Its mean represents the average forecasting performance, while its variance indicates the stability of forecasting performance. We find that our monitoring forecast beats the proposed forecast in both aspects. Our approach improves the mean prediction loss difference by 15%, while the variance falls by 46.3% because of its expected benefits. We emphasize two things here. First, our approach aims to attain a realistic level of out-of-sample prediction rather than compete with other algorithms that maximize ex-post performance metrics. Second, a robust monitoring machine can be easily added to other forecasting methods as a final step. In particular, forecasters should apply the robust monitoring machine if they evaluate performances relative to a benchmark.

We also confirm that the robust monitoring machine captures the well-known economic intuition in real-time: return predictability is greater in bad economic periods. The robust monitoring machine tends to favor the proposed forecast ex-ante in bad economic periods, which are characterized by low stock valuation and high macro uncertainty. Previously, this intuition looked clear only when researchers studied the whole sample period in hindsight.

This study contributes to two strands of the literature. First, a few recent works use a forward-looking approach to model selection or averaging; Zhu and Timmermann (2017), Gibbs and Vasnev (2018), and Granziera and Sekhposyan (2019) optimize weighting strategy conditional on the expected future performance of the prediction models. This approach represents a considerable departure from the prior literature with backward-looking approaches. We then fully scale up their forward-looking approach by searching the entire time-series model space as approximated by feature engineering (e.g., Kou et al. 2021), instead of considering only a few models possibly chosen by subjective intuitions. Second, stock return predictions via machine learning made

tremendous progress recently in academia and practice.[5] Prevailing research, however, concentrates on improving the accuracy of the prediction.[6] Instead, we exploit the black-box-like nature of machine learning to intentionally block intuition from the hypothetical future data (test sample) in out-of-sample simulations. That is, this study resolves the out-of-sample R^2-hacking problem by machine learning techniques instead of creating one. Unlike existing solutions for multiple testing, our approach does not compromise forecasting performance.

To the best of our knowledge, no prior works attempt to resolve the R^2-hacking problem by machine learning techniques. This gap creates the need for a new approach. This study offers the first attempt by proposing how to properly use existing machine learning algorithms to avoid the out-of-sample R^2-hacking problem instead of proposing a single machine learning algorithm that maximizes forecasting accuracy by multiple testing.

The remainder of the paper is organized as follows. Section Monitoring forecasts revisited revisits forecast monitoring in a simple framework. Section Robust monitoring explains the benefits of the robust monitoring forecast. Section Robust monitoring machine outlines our machine learning approach to robust monitoring. We describe its real-world example in Sect. Applications: robust monitoring for return-predictability and present the results in Sect. Empirical performance of monitoring forecasts. We discuss several important topics, such as how to interpret the results in conjunction with business cycles in Sect. Discussion and conclude in Sect. Conclusion.

Monitoring forecasts revisited

Monitoring forecasts can be viewed as an extreme case of combination forecasts with dynamic weights. If a researcher is unsure which forecast predicts conditionally best, it is better to combine multiple forecasts for variance reduction. However, selecting a single predictor can be better if an accurate signal for conditional performance is available. In this section, we characterize the condition on which a monitoring forecast can outperform individual forecasts in a simple model. We also include a few metrics and statistical tests to quantify monitoring performance.

A simple model of monitoring

Suppose that r is the target variable to predict and its unconditional mean is μ. There are two unbiased forecasts, $f^{(a)}$ and $f^{(b)}$; that is, $E[f^{(a)}] = E[f^{(b)}] = E[r] = \mu$, as follows.

$$r = \mu + e^{(y)} + e^{(s)}, \tag{1}$$

$$f^{(a)} = \mu + e^{(a)} + Se^{(s)}, \tag{2}$$

$$f^{(b)} = \mu + e^{(b)} + (1 - S)e^{(s)}. \tag{3}$$

[5] See (Goldstein et al. 2021; Kou et al. 2019), and (Abad-Segura and González-Zamar 2020) for recent advanced research in finance.

[6] See, for example, Gu et al. (2020), Feng et al. (2020), Heaton et al. (2017), and Freyberger et al. (2020) among many others.

The individual shocks $(e^{(y)}, e^{(s)}, e^{(a)}, e^{(b)})$ follow a normal distribution centered at zero, respectively. These shocks are uncorrelated to each other, and their standard deviations are $SD(e^{(y)}) = \sigma_y$, $SD(e^{(s)}) = \sigma_s$, and $SD(e^{(a)}) = SD(e^{(b)}) = \sigma_f$, respectively. The indicator variable S follows a Bernoulli distribution with probability p_s.

$$S = \begin{cases} 1 \text{ with probabiliy } p_s, \\ 0 \text{ with probabiliy } 1 - p_s. \end{cases} \tag{4}$$

Without loss of generality, we assume that the forecast $f^{(a)}$ is unconditionally more accurate than $f^{(b)}$ such that $p_s > \frac{1}{2}$. If $S = 1$, then the forecast $f^{(a)}$ is conditionally more accurate than $f^{(b)}$ ex-ante, while the forecast $f^{(b)}$ is if $S = 0$.

Now consider a monitoring forecast $f^{(m)}$ optimally switching between two forecasts $f^{(a)}$ and $f^{(b)}$ given a signal on S. We measure the accuracy of the signal, p_m, as

$$f^{(m)} = Mf^{(a)} + (1 - M)f^{(b)} \quad \text{where} \quad M = \begin{cases} S & \text{with probabiliy } p_m, \\ 1 - S & \text{with probabiliy } 1 - p_m. \end{cases} \tag{5}$$

M is a monitoring signal following a Bernoulli distribution. The higher p_m is, the more accurate the monitoring forecast is. For example, if $p_m = 1$, then the monitoring forecast always picks the more accurate forecast between $f^{(a)}$ and $f^{(b)}$. Next, we adopt a squared forecast error as a loss function:

$$L^{(a)} = (r - f^{(a)})^2, \;\; L^{(b)} = (r - f^{(b)})^2, \;\; \text{and} \;\; L^{(m)} = (r - f^{(m)})^2, \tag{6}$$

and define a loss difference between two forecasts as follows.

$$\Delta L^{(b,a)} = L^{(b)} - L^{(a)}, \;\; \Delta L^{(b,m)} = L^{(b)} - L^{(m)}, \;\; \text{and} \;\; \Delta L^{(a,m)} = L^{(a)} - L^{(m)}. \tag{7}$$

Then, we can easily derive the condition for which the monitoring forecast $f^{(m)}$ outperforms both individual forecasts on average. The monitoring forecast $f^{(m)}$ is unconditionally more accurate than $f^{(a)}$ (and so $f^{(b)}$) ex-ante if and only if

$$E[\Delta L^{(a,m)}] = E[\Delta L^{(b,m)}] - E[\Delta L^{(b,a)}] = (p_m - p_s)\sigma_s^2 > 0. \tag{8}$$

Condition (8) states that the monitoring forecast will outperform if a monitoring signal is accurate enough to dominate the accuracy advantage of an individual forecast over the other. For example, in case two forecasts are similar in accuracy $p_s = 1/2$, then it is not difficult to beat both forecasts by monitoring; we need only $p_m > 1/2$.

Out-of-sample R^2: a conventional metric for forecast performance

Campbell and Thompson (2008) suggests an out-of-sample R^2 metric to evaluate the performances of stock return forecasts. Suppose $f^{(i)}_{t|t-1}$ is a given forecast for the target r_t. Then, the out-of-sample R^2 for the period from $t = t_0$ to $t = t_1$ is

$$R^2_{OS} = 1 - \frac{\sum_{t=t_0}^{t_1}(r_t - f^{(i)}_{t|t-1})^2}{\sum_{t=t_0}^{t_1}(r_t - f^{(b)}_{t|t-1})^2}, \tag{9}$$

where $f_{t|t-1}^{(b)}$ is a benchmark forecast, which is the historical average of the past returns r_t, commonly used in the stock return predictability literature. The out-of-sample R^2 measures the reduction in mean squared prediction error for a forecast relative to the benchmark forecast. If R_{OS}^2 is positive, then the forecast $f_{t|t-1}^{(i)}$ outperforms the benchmark forecast in terms of the mean square prediction error metric.

Metrics and tests for monitoring performance

Monitoring forecasts solves a classification problem: which forecast performs better conditionally? Therefore, we can adopt metrics and statistical tests from the classification literature to evaluate monitoring performance. For example, we can define the event in which $f^{(a)}$ outperforms $f^{(b)}$; that is, $\Delta L^{(b,a)} > 0$, is *positive*, and *negative* otherwise. Then, the *sensitivity*, or true positive rate (TPR), refers to the empirical probability that the monitoring forecast $f^{(m)}$ equals $f^{(a)}$ when $\Delta L^{(b,a)} > 0$ (i.e., $f^{(a)}$ outperforms $f^{(b)}$). Likewise, *specificity*, or true negative rate (TNR), refers to the empirical probability that $f^{(m)}$ equals $f^{(b)}$ when $\Delta L^{(b,a)} < 0$. We also adopt other metrics, such as positive predicted value (PPV), negative predictive value (NPV), and accuracy (ACC), following their conventional definitions in a confusion matrix.

If the classifier is purely random, then both "sensitivity + specificity" and "PPV + NPV" should be one in the population. We compute their 95% confidence intervals as

$$TPR + TNR \pm 1.96 \times \sqrt{TPR \times (1 - TPR)/n_1 + TNR \times (1 - TNR)/n_2},$$

where n_1 and n_2 are the numbers of true positives and negatives in the data, respectively.

$$PPV + NPV \pm 1.96 \times \sqrt{PPV \times (1 - PPV)/n_3 + NPV \times (1 - NPV)/n_4}$$

where n_3 and n_4 are the numbers of predicted positives and negatives in the data, respectively. If these confidence intervals do not contain one, then we can conclude that a monitoring task is informative ($p_m > 1/2$) at the 5% significance level.

We adopt two formal tests of monitoring performance from the classification context: Fisher's Exact test and the Chi-square test for binary classifiers. The null hypothesis H_0 in both tests is that the true (predicted) positives and true (predicted) negatives are equally likely to be predicted as (true) positives. Therefore, low p-values of these tests are evidence that a monitoring task is informative.

Robust monitoring

Monitoring Forecast is an aggressive technique to maximize predictability in contrast to a *Combination Forecast*, which aims to reduce the variance of forecast errors. That is, monitoring forecast and combination forecast are traditional counterparts of boosting and bagging (i.e., bootstrap aggregating) in machine learning, respectively. However, monitoring the forecast, which is aggressive by nature, can produce a robust conservative predictor when it switches between a given proposed forecast and a benchmark. That is, monitoring forecast becomes a robust version of the originally proposed forecast, whatever it is. We briefly explain the intuition in the following sections.

Using a benchmark to make a forecast robust

Suppose a researcher wants to measure the forecasting ability of a proposed forecast $f^{(a)}$ relative to a benchmark forecast $f^{(b)}$ in the context of the previous section. Then, the loss difference, $\Delta L^{(b,a)} = L^{(b)} - L^{(a)}$, between the forecasts $f^{(a)}$ and $f^{(b)}$ is a measure of how much the forecast $f^{(a)}$ outperforms the benchmark forecast $f^{(b)}$. With a quadratic loss function, the expected loss difference $E(\Delta L^{(b,a)})$ is the expected reduction in mean squared error when we replace the benchmark $f^{(b)}$ by $f^{(a)}$.

Moreover, researchers can construct a monitoring forecast $f^{(m)}$ that switches between a proposed forecast $f^{(a)}$ and a benchmark forecast $f^{(b)}$. They may wonder if the monitoring forecast $f^{(m)}$ outperforms the proposed forecast $f^{(a)}$. Note that comparing the mean squared errors of the forecasts $f^{(a)}$ and $f^{(m)}$ is equivalent to comparing the expected loss differences $E(\Delta L^{(b,m)})$ and $E(\Delta L^{(b,a)})$ because of the following identity:

$$E[L^{(a)}] - E[L^{(m)}] = E[\Delta L^{(b,m)}] - E[\Delta L^{(b,a)}].$$

Here, the empirical loss difference ΔL is the main building block for evaluating relative forecasting performance. Its first moment is a difference in the mean squared errors, a key comparison metric in the forecasting literature. Researchers, therefore, compare their means $E(\Delta L^{(b,m)})$ and $E(\Delta L^{(b,a)})$ to see if the monitoring forecast $f^{(m)}$ outperforms the original forecast $f^{(a)}$ on average. However, what about their variances $Var(\Delta L^{(b,m)})$ and $Var(\Delta L^{(b,a)})$? Should researchers ignore or care about them? What do they even mean?

Those variances represent the uncertainty of how well the forecasts $f^{(a)}$ and $f^{(m)}$ perform at a given time relative to the benchmark forecast $f^{(b)}$. Yae (2018) adopts the idea of a tracking-error-volatility (TEV)-efficient frontier from the portfolio optimization literature and argues that risk-averse researchers should care about such variances if they evaluate performance relative to the benchmark forecast.[7] The idea of relative performance is common in investments. An investor who wants to outperform the market would consider deviating from the market as taking risks for higher returns. The more the portfolio deviates from the market, the greater the downside risk (and upside opportunity) relative to the market performance.

Nevertheless, conventional forecasting performance metrics focus only on the mean performances while ignoring information in the variance of performances. Such second-moment information is used only in some formal tests of relative forecast performance such as Diebold and Mariano (2002).[8] This first-moment-oriented practice should look alarming to financial economists because it ignores risk, which is the core of investment performance evaluation. Additionally, the variance of relative performance is actually common in economics and finance.[9] For example, in "Keeping-up-with-the-Joneses"

[7] Roll (1992) defines tracking error volatility as a square root of the sample second moment of differences in a portfolio and benchmark return. The TEV-efficient frontier shows a trade-off between relative risk premium and relative risk to the benchmark, while the standard efficient frontier in the portfolio theory is a special case where the benchmark is the risk-free asset.

[8] Note this second moment is not the variance of forecast errors but that of differences in squared forecast errors, defined relative to the choice of benchmark. That is, the variance of relative performance is related to the fourth, not the second, moment of forecast errors.

[9] However, unlike the first moment, the second (central) moment comparison requires caution with existence of the benchmark forecast:

$$Var(L^{(a)}) - Var(L^{(m)}) \neq Var(\Delta L^{(b,a)}) - Var(\Delta L^{(b,m)}).$$

preferences, an agent's utility is determined relative to others' consumption level (Abel 1990, Gali 1994). Similarly, stock market movements do not compensate or penalize a mutual fund manager whose official benchmark is the market portfolio.

A *Robust monitoring forecast* is "a monitoring forecast switching between a proposed forecast and a benchmark." It has a built-in shrinkage effect when the forecast is evaluated relative to a benchmark in terms of $\Delta L^{(b,m)}$. The monitoring technique in this context makes a newly proposed forecast robust, and exploits shifts in the conditional predictability between two forecasts. Its mechanism is simple. If monitoring is informative, then the monitoring forecast will become a benchmark forecast $f^{(m)} = f^{(b)}$ when the other forecast $f^{(a)}$ is unlikely to outperform the benchmark. Whenever $f^{(m)} = f^{(b)}$, the loss difference $\Delta L^{(b,m)}$ becomes exactly zero, and the total variance of the loss difference becomes lower than that of the proposed forecast $f^{(a)}$. In the monitoring model of the previous section, the law of total variance implies

$$\begin{aligned} Var(\Delta L^{(b,m)}) &= E[Var(\Delta L^{(b,m)}|M)] + Var[E(\Delta L^{(b,m)}|M)] \\ &= p_a Var(\Delta L^{(b,a)}) + p_a(1-p_a)\{E(\Delta L^{(b,a)})\}^2, \end{aligned}$$

where $p_a \equiv Prob[f^{(m)} = f^{(a)}] = p_m p_s + (1-p_m)(1-p_s)$ denotes the unconditional probability that the monitoring forecast deviates from the benchmark $f^{(m)} = f^{(a)}$. The second term is negligible unless the proposed forecast $f^{(a)}$ outperforms the benchmark significantly and persistently, which implies the benchmark is improperly chosen as a straw man. Therefore, we can approximate the ratio of variance of loss differences $\Delta L^{(b,m)}$ as p_a:

$$\frac{Var(\Delta L^{(b,m)})}{Var(\Delta L^{(b,a)})} \approx p_a \leq p_s < 1 \quad \text{and} \quad p_a \in [1-p_s, p_s]. \tag{10}$$

The variance ratio in the left-hand-side is always lower than one by construction. Therefore, monitoring with a benchmark forecast always lowers the variance of loss difference $\Delta L^{(b,m)}$, and the monitoring forecast will produce more statistically significant evidence on superior forecasting performance. This is a hidden benefit of forecast monitoring.

However, this benefit is never a free lunch. The following trade-off between accuracy gain and variance reduction in loss difference exists[10]:

$$\underbrace{\left[\frac{2p_s - 1}{\sigma_s^2}\right]}_{\text{Constant (+)}} \cdot \underbrace{\left[E(\Delta L^{(b,m)}) - E(\Delta L^{(b,a)})\right]}_{\text{Increase in average accuracy}} + \underbrace{(1-p_a)}_{\substack{\text{Reduction} \\ \text{in variance} \\ \text{of loss diff.}}} = \underbrace{2p_s(1-p_s)}_{\text{Constant (+)}}. \tag{11}$$

We derive this result by eliminating p_m by combining $p_a = p_m p_s + (1-p_m)(1-p_s)$ and Eq. (8). Informative monitoring (i.e., high p_m) improves the accuracy of the monitoring forecast but decreases the reduction in loss-difference variance, as follows.

[10] This trade-off relationship is from Yae (2018).

$$\frac{\partial(1-p_a)}{\partial p_m} = 1 - 2p_s < 0 \quad \text{if} \quad p_s > \frac{1}{2}. \tag{12}$$

Note that this trade-off is beyond the bias-variance trade-off in statistical estimators or machine learning algorithms. The bias-variance trade-off aims to maximize the average forecasting performance. Equation (11) then represents a trade-off between the average forecasting performance and its variance. This new kind of trade is about the mean and variance of squared forecast errors, which correspond to the second and fourth moments of forecast errors, while the bias-variance trade-off is about the first and second moments of forecast errors. In a numerical example with $p_s = p_m = 3/4$, we can still expect a 37.5% reduction in variance of loss-difference with the same accuracy as the originally proposed forecast in terms of the mean squared error.

The definition of variance of loss difference depends on the choice of benchmark forecast. In many applications, the benchmark is not subjective. For example, the classical view in the stock market is that price follows a random walk, with absolutely no predictability. In other words, the market risk premium is unconditionally constant and investors have no conditioning variables to predict it. If researchers want to show the existence of a successful predictor or predictability of market risk premium, then they needs to set up a constant market risk premium as a null hypothesis and use a historical average stock market return as a benchmark forecast.[11]

Metrics for robust monitoring performance

We adopt two metrics for robust monitoring performance from Yae (2018). The metrics are analogous to investment performance measures: the risk premium as a raw metric and alpha as a risk-adjusted metric.[12] Here, we use concise notations for performance metric inputs:

$$d_m \triangleq \Delta L^{(b,m)} \quad \text{and} \quad d_a \triangleq \Delta L^{(b,a)}.$$

Monitoring Risk Premium Suppose a researcher chooses a forecast $f^{(a)}$ when considering a benchmark forecast $f^{(b)}$ as a reference point. Then, $E[d_a]$ can represent the average forecasting performance as the expected reduction in mean squared forecast errors relative to the benchmark. If the researcher chooses $f^{(b)}$ instead, then she obtains only $E[d_b]$, which is zero by definition. We interpret $E[d_a]$ as a kind of premium in the forecasting context. As an analogy, if an investor chooses the stock market portfolio over her benchmark risk-free asset, then the expected return difference between these two is called a market risk premium. The investor deviates from the benchmark risk-free asset in hope of earning the premium. Likewise, a researcher deviates from the benchmark forecast $f^{(b)}$ to $f^{(a)}$ in hope of earning $E[d_a]$. She can earn $E[d_m]$ instead by switching between $f^{(a)}$ and $f^{(b)}$ as a robust monitoring forecast $f^{(m)}$. Then, $E[d_m]$ scaled by $E[d_a]$ is called the *monitoring risk premium*.

[11] Sect. 5 shows an empirical example.

[12] We do not consider a utility-function-based measure such as a certainty equivalent although, for example, smooth ambiguity preference (Klibanoff et al. 2005) can internalize $Var(\Delta L^{(b,m)})$ and $Var(\Delta L^{(b,a)})$.

$$\text{(Monitoring Risk Premium):}\quad RP_m = \frac{E[d_m]}{E[d_a]} = \frac{R^2_{OS,m}}{R^2_{OS,a}}, \tag{13}$$

where $R^2_{OS,a}$ and $R^2_{OS,m}$ are the out-of-sample R^2 for the proposed forecast $f^{(a)}$ and the robust monitoring forecast $f^{(m)}$ that switches between $f^{(a)}$ and the benchmark forecast $f^{(b)}$, respectively. If monitoring is sufficiently informative as in Condition (8), then RP_m can be larger than one.

Monitoring Alpha Monitoring risk premium measures average forecasting performance but ignores any risk adjustment. By contrast, imagine investors who use CAPM alpha (or alpha from a multi-factor model) as a risk-adjusted investment performance metric. Similarly, we define the monitoring alpha as

$$\text{(Monitoring Alpha):}\quad \alpha_m = \frac{E[d_m] - E[d_{m^*}]}{E[d_a]}, \tag{14}$$

where m^* is the uninformed monitoring forecast that shifts randomly between the forecasts $f^{(a)}$ and $f^{(b)}$. It is a random strategy whose mixing probability is set so that its $E[d^2_{m^*}]$ equals $E[d^2_m]$. Thus, $E[d_{m^*}]$ represents accuracy gain by randomly shifting between the forecasts, and the real gain by informative monitoring should not include $E[d_{m^*}]$. For example, if $\alpha_m = 0.4$, then informative monitoring adds 40% of relative accuracy of the proposed forecast on top of the benefit from random switching. Alternatively, we can also express the monitoring alpha as[13]

$$\alpha_m = \frac{E[d_m]}{E[d_a]} - \frac{E[d^2_m]}{E[d^2_a]}. \tag{15}$$

The first term is the total accuracy gain through monitoring—scaled by the accuracy gain through the forecast $f^{(a)}$; that is, the monitoring risk premium, while the second term adjusts the relative risk increased by the monitoring procedure. When the unpredictable component is large in its scale σ_y, we have $(E[d_m])^2 \ll Var[d_m]$, so $E[d^2_m] \approx Var[d_m]$. Therefore, the monitoring alpha is analogous to a utility level of mean-variance preference. Note that the monitoring alpha of the two forecasts $f^{(a)}$ and $f^{(b)}$ are zero by definition.

Robust monitoring machine

Data snooping is a common issue in empirical research. The problem arises when a researcher reports only the best model or statistically significant variable after numerous failed trials. The same problem can also appear when numerous researchers try only one model or variable, but only a few researchers can successfully publish their results, which is dictated by luck, as shown in Chordia et al. (2017). This fundamental problem of empirical research is difficult to avoid and persists even in the forecasting context. Robust monitoring is not an exception.

[13] It is easy to show $E[d_{m^*}]/E[d^2_{m^*}]$ is invariant and so $E[d_{m^*}]$ is linearly proportional to $E[d^2_{m^*}]$.

Out-of-sample R2-hacking problem

Data snooping (or p-hacking) comes from two root causes: data and models (algorithms). First, researchers face infinite combinations of choices of variables, sample periods, and training-evaluation sample splits. Second, they must also choose a model along with tuning parameters, algorithms, and estimation methods. Analyzing the whole data-model space exceeds an individual researcher's cognitive ability. Therefore, they end up selecting one or a few combinations arbitrarily or deliberately, which might even worsen data snooping issues.

The p-hacking problem regarding variable choices is well known in the in-sample fitting context, but the same problem also exists in out-of-sample analysis regarding both variable and model choices. For example, in the return predictability literature, out-of-sample R^2 is mainly used as a forecasting performance metric. Nonetheless, out-of-sample R^2 (or any other cross-validation metric) still faces the multiple testing problem and look-ahead bias. For example, some early papers, such as the one by Pesaran and Timmermann (1995), report that predictability is time-varying and conditional on other variables. That is, the predictability is stronger in recessions than during expansions. Then, many follow-up papers internalize such empirical facts in a specific time-series model to further improve out-of-sample predictability, ex-post.

The pitfall of this practice is that researchers select such successful models and variables based on their intuition, which did not exist in the sample period of the back tests. The intuition is formed by empirical knowledge only available now but unavailable at the beginning of the out-of-sample test period, such as information that the predictability is stronger in recession than expansions. Therefore, applying such models and variables since the beginning of the out-of-sample test period is highly unlikely for the real forecasters at that time due to lack of prior evidence. That is, all sophisticated models inspired by such ex-post intuition are subject to this unintended look-ahead bias. The best option for a forecaster at that time, if feasible, was to compare all possible models and variables to find the best one (or best combination) ex-ante since the beginning of the out-of-sample testing period while repeating the process sequentially.

A machine learning solution

Machine learning algorithms are often criticized because of their black-box-like nature. Despite their superior prediction ability, they make researchers blind to hidden mechanisms by blurring economic intuition. Here, we focus on the bright side of the black-box-like nature and transform this criticism into a crucial device in our study. We make our solution for the robust monitoring problem intentionally blind to any intuition based on the information in the evaluation (test) sample period, as Yae (Forthcoming) suggests. Therefore, the goal of our approach is to confirm the existence of useful information for monitoring in the real-time data rather than maximize the average forecasting performance ex-post. To achieve this goal, we implement the following three steps.

Robust Monitoring Machine

1. Choose a set of conditioning variables for monitoring, independent of information from the evaluation sample period.

2. Approximate the entire time-series model space of the conditioning variables with feature engineering.
3. Apply a combination of ensemble machine learning algorithms instead of choosing what works best ex-post after trying many.

Each step is designed to exclude the out-of-sample R^2-hacking associated with selecting data, models, and optimization algorithms. Therefore, how to implement the first two steps are as important as the last step. We emphasize that this three-step approach is beyond a well-known ensemble technique which is only the step 3 here.

We call this machine learning approach the *robust monitoring machine* and implement each step as follows. First, we make the most parsimonious choice of conditioning variables rather than the most universal. In monitoring, a researcher needs to have at least one forecast target and two competing forecasts. We use a single time-series loss difference between two competing individual forecasts, $\Delta L_t^{(b,a)}$, to predict its next sign. The other extreme is to consider the entire information space, yet it is difficult to fathom, collect, or even approximate. Furthermore, many data sources are private or highly costly and are thus out of reach of some researchers. As the entire information space is neither known nor accessible, random sampling from it is also impossible. Second, we expand a single time-series of a conditioning variable $\Delta L_t^{(b,a)}$ into hundreds of time-series features as building blocks to approximate the entire time-series model space. Finally, we combine multiple ensemble machine learning algorithms. We rely on machine learning algorithms to handle the complexity of the feature set, but avoid cherry-picking an algorithm that appears to work best ex-post.

Applications: robust monitoring for return-predictability

We apply our robust monitoring machine to the US stock market prediction problem. Section Data, benchmark, and proposed forecasts explains the data source and our choice of benchmark and proposed forecasts. Section Robust monitoring machine in action describes our three-step approach in detail.

Data, benchmark, and proposed forecasts

The random-walk hypothesis is the traditional view of the stock market. If investors are rational and fully utilize public information to price stocks, then such public information should not predict future stock returns. Following the literature on return predictability, we set the target variable to predict, as monthly stock market excess returns: continuously compounded returns $r_{sp,t+1}$ on the S &P 500 index, including dividends, in excess of the risk-free rate $r_{f,t}$ implied by the Treasury bill rate.[14] Henceforth, we call the target variable simply "return."

$$\textit{Target Variable: } r_{t+1} = r_{sp,t+1} - r_{f,t}. \tag{16}$$

Suppose the random-walk hypothesis is true. Then, econometricians will find that no publicly available variables are correlated with the subsequent return r_{t+1} or its

[14] Since the risk-free rate is known at the time of forecast, predicting raw returns is informationally equivalent to predicting excess returns.

conditional expectation $E[r_{t+1}|\mathcal{F}_t]$, where $\mathcal{F}_t$ denotes the information set of econometricians at time t. The expected return, therefore, should look like a constant $E[r_{t+1}|\mathcal{F}_t] = E[r_{t+1}]$, whether it truly is or not. The natural benchmark forecast under the random-walk hypothesis as a null will be an estimate for the unconditional expected return $E[r_{t+1}]$. That is, the historical average of the past returns is the benchmark forecast.

$$\textit{Benchmark Forecast: } f^{(b)}_{t+1|t} = \frac{1}{t}\sum_{\tau=1}^{t} r_\tau. \tag{17}$$

The idea of the random-walk hypothesis is theoretically appealing and relates to the well-known efficient market hypothesis. However, it was mostly difficult to reject the hypothesis early on because of the insufficient sample size. Later researchers, however, began to find some statistical evidence on a few variables predicting stock market returns based on in-sample regression analysis.[15] However, Goyal and Welch (2008) examine the out-of-sample performance of the real-time individual OLS regression estimators of 14 variables $x_{i,t}$ and find that none of them have out-of-sample predictability. From the following 14 individual predictive regressions:

$$r_{t+1} = \alpha_i + \beta_i x_{i,t} + \epsilon_{i,t+1} \quad \text{for} \quad i = 1, ..., 14. \tag{18}$$

Goyal and Welch (2008) and Rapach et al. (2010) define the real-time individual forecasts as

$$f^{(i)}_{t+1|t} = \hat{\alpha}_{i,t} + \hat{\beta}_{i,t} x_{i,t}, \tag{19}$$

where $\hat{\alpha}_{i,t}$ and $\hat{\beta}_{i,t}$ are the OLS coefficient estimates of α_i and β_i, respectively, using information up to time t (expanding window), consistent with Goyal and Welch (2008) and Rapach et al. (2010). Later, following Hendry and Clements (2004), Timmermann (2006), and many others, Rapach et al. (2010) construct the equal-weighted average of these 14 individual forecasts as follows and show it outperforms the benchmark forecast (17) in out-of-sample prediction.

$$\textit{Proposed Forecast: } f^{(a)}_{t+1|t} = \frac{1}{14}\sum_{i=1}^{14} f^{(i)}_{t+1|t} \tag{20}$$

We use this equal-weight combination forecast as our proposed forecast in monitoring.[16] Then, we will optimally switch our choice on forecast between the benchmark forecast in (17) and the proposed forecast in (20).

We obtain the monthly data for the returns and predictor variables from Amit Goyal's website.[17] The sample period is from January 1927 to December 2017. Note that the

[15] It is worth noting that the rejection of the random-walk hypothesis does not necessarily mean the stock market is inefficient in information processing or investors are irrationally inattentive. It simply means the expected return is time-varying and correlated with some publicly available variables because, roughly speaking, investors' risk tolerance is time-varying. For example, in recession investors become less risk-tolerant because of their reduced income and wealth. They avoid investing in stocks even if they know the expected return is higher than in the boom periods.

[16] Alternatively, we can optimize the combination weights or construct a new forecast utilizing nonlinearity and interactions between 14 variables. However, our goal is to build a robust and general approach that can be applied to forecasts from human forecasters or different models/algorithms as well.

[17] The data used in this paper can be found at http://www.hec.unil.ch/agoyal/ along with detailed descriptions.

combination forecast in our analysis differs slightly from that of Rapach et al. (2010) because their training sample starts in 1947. However, this is barely an issue because we do not attempt to criticize their combination forecast. On the contrary, we need only to demonstrate our idea using some forecast that overall outperforms the benchmark forecast but its performance varies. The 14 predictor variables in our example include dividend-price ratio (dp), dividend yield (dy), earning-price ratio (ep), dividend-payout ratio (de), equity risk premium volatility (rvol), book-to-market ratio (bm), net equity expansion (ntis), treasury bill rate (tbl), long-term yield (lty), long-term return (ltr), term spread (tms), default yield spread (dfy), default return spread (dfr), and inflation (infl). See Appendix 1 for detailed descriptions.

Robust monitoring machine in action

This section explains how we implement the robust monitoring machine in detail.

Labeling for Binary Classification We define a new dependent variable y_t as follows to convert the monitoring problem to a binary classification.

$$y_t = \begin{cases} 1, \text{ if } \Delta L_t^{(b,a)} > 0, \\ 0, \text{ otherwise,} \end{cases} \tag{21}$$

where $\Delta L_t^{(b,a)} = (r_t - f_{t|t-1}^{(b)})^2 - (r_t - f_{t|t-1}^{(a)})^2$ is the difference in squared forecast errors as a loss function. Target variable r_t, benchmark forecast $f_{t|t-1}^{(b)}$, and proposed forecast $f_{t|t-1}^{(a)}$ are as defined in Sect. Data, benchmark, and proposed forecasts. This binary time-series variable y_t, as a label, indicates which forecast outperform the other, ex-post. Note the resulting dependent variable y_t is identical whether the definition of $\Delta L_t^{(b,a)}$ is based on L^2-norm or L^1-norm.

Features for Monitoring As we discussed in Sect. A machine learning solution, we include no arbitrary conditioning variables in the monitoring task. Instead, we use a single time-series of $\Delta L_{t-1}^{(b,a)}$ to predict y_t. Following (Christ et al. 2018), we perform feature engineering. We transform the past 60 months of $\Delta L_t^{(b,a)}$ into 441 time-series model features Z_{t-1}.[18] These features include the entire characteristics of the time-series such as the number of peaks, average, maximal value, autocorrelation, and linear trend.[19]

$$g : X_{t-1} \to Z_{t-1} \text{ where } X_{t-1} = \{\Delta L_{t-1}^{(b,a)}, \Delta L_{t-2}^{(b,a)}, ..., \Delta L_{t-60}^{(b,a)}\}.$$

The features for monitoring Z_{t-1} offer building blocks to approximate the entire space of time-series models that predict y_t.

Three Ensemble Decision-Tree Algorithms To handle hundreds of features, we use the three flagship ensemble decision-tree algorithms to predict y_t: the random forest, extremely randomized trees, and gradient boosting. They can effectively accommodate nonlinearity and interaction in features. They also avoid overfitting problems in traditional logistic regressions by combining the forecasts from many small trees into a single forecast. We summarize the technical differences of these three algorithms in Appendix 3.

[18] Using the past sixty month is a standard practice in finance literature due to changing nature of the market: for example, CAPM beta estimation.

[19] See TSFRESH package in Python.

Training and Tuning Parameters To predict y_t, we train the tree using input-label pairs $(y_{t-\tau}, Z_{t-\tau-1})$ for $\tau = 1, 2, ..., 120$ (ten years of monthly data). As feature engineering requires the past 60 months of data, we need 15 years of data to predict y_t. Note that using rolling-window is a natural choice since the idea of monitoring is based on time-varying performance. We choose a tuning parameter that maximizes the ROC-AUC statistic measure which is the Area Under The Curve (AUC) of Receiver Operating Characteristics (ROC) curve. In simple terms, this procedure maximizes TPRs relative to false positive rates. We split these 120 sample pairs into three of 40 sample pairs D_1, D_2, and D_3 chronologically. Then, we train trees using D_1 and test on D_2. Again, we train the tree using D_1 and D_2 and test on D_3. We choose the optimal tuning parameters based on these two test results (i.e., validation sample). The out-of-sample forecast starts from January 1947, following (Goyal and Welch 2008).

Ensemble At time t, the three different algorithms will produces their best guesses on the probability of $y_{t+1} = 1$; say, $Prob[y_{t+1} = 1|A_i, \mathcal{F}_t]$ for $i = 1, 2, 3$, where A_i is an algorithm and $\mathcal{F}_t$ is an information set. Then, we compute the equal-weight average of such probabilities relying on the wisdom of crowds in the algorithm domain.

$$Prob[y_{t+1} = 1|\mathcal{F}_t] = \frac{1}{3}\sum_{i}^{3} Prob[y_{t+1} = 1|A_i, \mathcal{F}_t], \tag{22}$$

The following rule will determine our monitoring forecast of y_{t+1} at time t:

$$f^{(m)}_{t+1|t} = \begin{cases} f^{(a)}_{t+1|t}, \text{ if } Prob[y_{t+1} = 1|\mathcal{F}_t] > \frac{1}{2}, \\ f^{(b)}_{t+1|t}, \text{ otherwise.} \end{cases} \tag{23}$$

Empirical performance of monitoring forecasts

Performance of the proposed forecast

The first row of Table 1 shows the performance of the proposed forecast $f^{(a)}$ in Eq. (20). The out-of-sample R^2 of the proposed forecast is 0.5% when the evaluation sample starts in 1947. The performance metric, however, is unstable and sensitive to the sample split date. When we gradually change the date from 1947 to 2007, the out-of-sample R^2 decreases and even becomes negative since 1987. The proposed forecast performs poorly, showing no predictability ($R^2_{OS} = -0.24\%$) when the evaluation sample starts from 2007 and ends in 2017.

Figure 1 Panel (a) also confirms that the performance of the proposed forecast $f^{(a)}$ significantly varies over time, with a downward trend. The time-series in the plot is the 60-month trailing moving average of the loss difference between the benchmark and the combination forecasts $\Delta L^{b,a}_t$, which measures the performance of the proposed forecast relative to the benchmark. The plot shows two deep negative values, implying that the proposed forecast often greatly underperforms the benchmark, especially in recent periods. Such unstable and deteriorating performance can be a serious concern to investors, although the proposed forecast overall outperforms the benchmark during the testing period 1947–2007.

Table 1 Out-of-sample $\mathbf{R^2}$ (%) of Traditional Combination Forecasts

	First year in evaluation sample						
	1947	**1957**	**1967**	**1977**	**1987**	**1997**	**2007**
Proposed forecast: $f^{(a)}$	0.50	0.37	0.36	0.14	− 0.09	− 0.10	− 0.24
Alternative combination forecasts							
Median	0.40	0.37	0.38	0.21	0.09	0.08	0.04
DMSFE (60 months, $\delta = 1.0$)	0.50	0.37	0.37	0.15	− 0.08	− 0.09	− 0.24
DMSFE (24 months, $\delta = 1.0$)	0.49	0.36	0.37	0.14	− 0.04	− 0.03	− 0.19
DMSFE (12 months, $\delta = 1.0$)	0.56	0.43	0.42	0.18	− 0.03	− 0.00	− 0.14
DMSFE (1 month, $\delta = 1.0$)	1.17	1.09	1.18	1.13	− 0.31	− 0.34	− 1.26
DMSFE (60 months, $\delta = 0.5$)	0.57	0.45	0.43	0.14	− 0.08	− 0.01	− 0.08
DMSFE (24 months, $\delta = 0.5$)	0.57	0.45	0.43	0.14	− 0.08	− 0.01	− 0.08
DMSFE (12 months, $\delta = 0.5$)	0.57	0.45	0.43	0.14	− 0.08	− 0.01	− 0.08

This table reports the forecasting performance of the proposed forecast and its variations, in terms of out-of-sample R^2 (%). The main proposed forecast is a equal-weighted combination forecast from Rapach et al. (2010). Its variational forms include a combination as a median and combinations using discounted mean square forecast error (DMSFE) from Stock and Watson (2004). We exploredifferent holdout windows of 1, 12, 24 and 60 months and discount factors of 0.5 and 1. Each column corresponds to a different sample split year from 1947 to 2007. All evaluation period ends in December 2017

The unstable performance of the proposed forecast is a fundamental problem of a class of combination forecasts, as Rapach et al. (2010) suggests. We explore different weighting schemes in combination forecasts to demonstrate the severity of the problem rather than propose improved weighting schemes. First, we use the median of 14 forecasts instead of their mean. Second, following (Stock and Watson 2004) Equation (4), we compute the combination weights of 14 individual forecasts based on their Discounted Mean Squared Forecast Errors (DMSFEs) with two tuning parameters: 1) the trailing sample period and 2) discount factor δ.

However, the unstable pattern of performance persists in all cases of different combination weights. For example, the out-of-sample R^2 can rise up to 1.17% with weights proportional to the DMSFE (up to the past one month of data with a discount factor $\delta = 1$). Yet the out-of-sample R^2 drops to -1.26% when the evaluation sample starts from 2007 and ends in 2017. Here, suppose we focus only on the average performance metric R^2_{OS}=1.17% and argue that we find a superior forecast. If we really do so, then this is cherry-picking, an example of out-of-sample R^2-hacking. Furthermore, it comes with a high price tag. This cherry-picked forecast has the worst performance of all forecasts over the recent decade. This unstable performance is a red flag that implies cherry-picking practices. To tackle this problem, we propose a machine learning solution in the next section.

Robust monitoring performance

We first evaluate how well our robust monitoring machine can conditionally choose between $f^{(a)}$ and $f^{(b)}$. The confusion matrix in Table 2 analyzes the monitoring performance as an out-of-sample classification problem without any look-ahead biases. Sensitivity and specificity are 57.0% and 56.8%, respectively, both higher than 50%. The 95% analytic confidence interval of the sum of sensitivity and specificity is (1.07, 1.21), above one, implying informative monitoring. Note that the sum of sensitivity and specificity

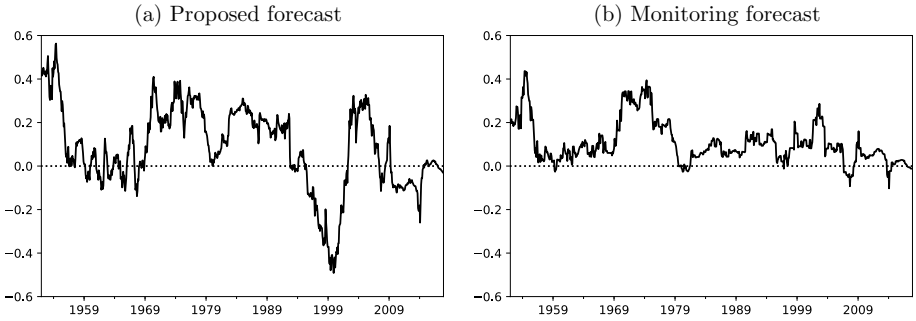

Fig. 1 Time-Varying Performances of the Proposed and the Monitoring Forecasts This figure plots the performances of the proposed and the monitoring forecasts relative to the benchmark. Positive values mean a forecast outperforms the benchmark. The horizontal axis is year while the vertical axis is the 60-month trailing moving average of difference in squared forecast error, multiplied by 10^4

should be one if a classifier is uninformative and purely random. Similarly, The PPV and NPV are 55.5% and 58.3%, respectively, and both are greater than 50%. The 95% analytic confidence interval of the sum of PPV and NPV is (1.07, 1.21), which is also above one. Furthermore, the p-values of Fisher's exact test and Chi-square test are both less than 0.01%. Our robust monitoring machine is informative in the out-of-sample prediction context.

Figure 2 visualizes two performance metrics for robust monitoring. The monitoring risk premium RP_m is the ratio of $E[d_m]$ to $E[d_a]$. The metric RP_m is larger than one since $E[d_m] > E[d_a]$, and therefore the monitoring forecast $f^{(m)}$ outperforms the proposed forecast $f^{(a)}$ in term of average out-of-sample R^2. On the other hand, the monitoring alpha α_m is the ratio of $E[d_m] - E[d_{m^*}]$ to $E[d_a]$. This metric adjusts the risk in forecasts with respect to the benchmark forecast $f^{(b)}$. Note that $E[d_{m^*}]$ represents the baseline performance of an uninformative monitoring forecast. Thus, the monitoring alpha is the net gain of performance by our robust monitoring machine, in addition to the risk reduction by robust monitoring described in Eqs. (10) and (11). The monitoring premium and alpha of our robust monitoring machine are 1.15 and 0.61, respectively, as Table 3 reports. Their Bootstrap p-values are lower than 5%. Robust monitoring machine truly predicts which forecast performs better. Our monitoring performance metrics are consistent with the Diebold-Marino (DM) test results. The p-values of the DM test for the robust monitoring machine is 0.4% while 6.8% for the proposed forecast. The robust monitoring machine shows stronger statistical significance (lower p-values) as $E[d_m] > E[d_a]$ though $Var[d_m] < Var[d_a]$. The monitoring forecast boosts the average loss difference by 15.1% but reduces its variance by 46.3%, which is a typical benefit of robust monitoring.

Figure 1 Panel (b) repeats Panel (a) but for the robust monitoring machine, showing the 60-month trailing average of $\Delta L_t^{(b,m)}$. This forecasting performance metric still fluctuates but rarely drops below zero. Its variation is much lower relative to the proposed forecast in Figure Panel (a). Table 4 shows the out-of-sample R^2 with different sample split dates. Unlike the proposed forecast $f^{(a)}$, the out-of-sample R^2 of the robust monitoring forecast $f^{(m)}$ never becomes negative. For the full sample period starting in 1947, the out-of-sample R^2 is 0.57% for the robust monitoring forecast and 0.50% for the proposed forecast. Our robust monitoring machine increases average

Table 2 Monitoring performance measures

	True positive	**True negative**	
Panel A: confusion matrix			
Predicted positive	236	189	PPV = 55.5%
Predicted negative	178	249	NPV = 58.3%
	TPR = 57.0% (Sensitivity)	TNR = 56.8% (Specificity)	Accuracy = 57.0%
	Estimate	**95% C.I.**	***P*-value**
Panel B: classification performance tests			
TPR + TNR	1.14	(1.07, 1.21)	
PPV + NPV	1.14	(1.07, 1.21)	
Fisher's exact test			6.84×10^{-5}
Chi-square test			5.29×10^{-5}

This table evaluates the performance of out-of-sample binary classification (without look-ahead bias) by robust monitoring machine for the evaluation period from 01/1947 to 01/2017. Panel A shows a confusion matrix based on predicted binary outcomes: positive (the proposed forecast outperforms the benchmark) or negative (otherwise). Sensitivity, or True Positive Rate (TPR), is calculated as the number of true positives divided by the number of real positives. Specificity, or True Negative Rate (TNR), is computed as the number of true negatives over the number of real negatives. PPV stands for Positive Predicted Value, calculated as the number of true positives divided by the number of predicted positives. NPV, or Negative Predictive Value, is computed as the number of true negatives divided by the number of predicted negatives. Accuracy is equal to the number of correctly predicted as a fraction of the total number of the sample observations. Panel B reports the 95% confident intervals for 'Sensitivity + Specificity' and 'PPV + NPV'

Table 3 Robust Monitoring Performance Measures

Forecast		**Monitoring Risk Premium**	**Monitoring Alpha**	$E[d_i]$ $\times 10^{-6}$	$Var[d_i]$ $\times 10^{-8}$	**Diebold-Mariano Statistic**
Robust	Estimate	1.15	0.61	9.81	0.99	2.87
Monitoring (m)	(*p*-value)	(0.017)	(0.021)			(0.004)
Proposed (a)	Estimate	1	0	8.52	1.86	1.82
		(by definition)	(by definition)			(0.068)
Benchmark (b)	By definition	0	0	0	0	N/A

This table reports monitoring performance measures: monitoring risk premium and monitoring alpha. $E[d_i]$ and $Var[d_i]$ for $i \in \{a, m, b\}$ denote the expected value and variance of loss difference, which represent average forecasting performance and its relative risk, respectively. Diebold-Mariano tests are to show if a forecast has significantly different predictability from a benchmark. The (two-sided) p-values in parentheses are calculated via bootstrapping for Monitoring Alpha and Monitoring Risk Premium. All forecasts are free from look-ahead bias

performance by 14% in terms of out-of-sample R^2 while reducing the variation of conditional performance. A shrinkage forecast, which is a simple average of $f^{(a)}$ and $f^{(b)}$, fails to improve the performance. We include a few alternative approaches to robust monitoring for comparison. They all fail to improve the proposed forecast; that is, the superior performance of the robust monitoring machine is not easy to achieve.

Certainty equivalent return

Following (Campbell and Thompson 2008) and (Ferreira and Santa-Clara 2011), we compute the certainty equivalent return (CER) for an investor with a mean-variance preference who monthly allocates her capital across equities and risk-free bills using market return forecasts. At the end of month t, the investor optimally allocates the share w_t of the portfolio to a market index fund and the remaining share $1 - w_t$ to a risk-free

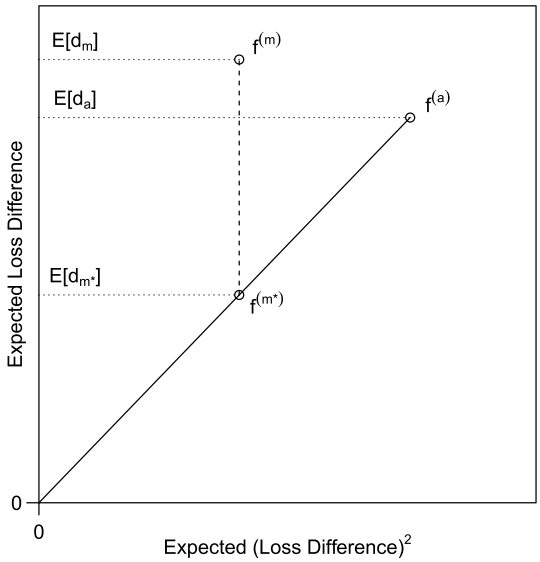

Fig. 2 Risk and Returns in Forecasting Performance This figure visualizes forecasts in the plot of risk and returns. The vertical axis is the the expected loss difference that represents the average forecasting performance relative to the benchmark. The horizontal axis is the expected squared loss difference as risk in forecasting performance. The benchmark forecast, by definition, is located at the origin. The proposed forecast $f^{(a)}$ and the monitoring forecast $f^{(m)}$ are also located in the plot according to their performance and risk metrics

bill. The investor holds the position until the end of month $t+1$ and repeats her asset allocation task every month. Then, we can compute the optimal share w_t as

$$w_t = \frac{1}{\gamma} \frac{E[r_{t+1}|\mathcal{F}_t]}{Var[r_{t+1}|\mathcal{F}_t]}, \tag{24}$$

where r_{t+1} is the excess return (raw return less the risk-free rate) and $E[r_{t+1}|\mathcal{F}_t]$ and $Var[r_{t+1}|\mathcal{F}_t]$ are its conditional mean and variance, respectively. Following (Campbell and Thompson 2008), we assume the followings. First, the investor replaces the conditional mean and variance by a given return forecast and simple variance estimator from the past five-year monthly returns. Second, we restrict the share proportion w_t to lie between 0 and 1.5. Third, we assume the relative risk coefficient γ to be 5. Then, the CER measure of the portfolio is

$$CER = \hat{\mu}_p - \frac{1}{2}\gamma\hat{\sigma}_p^2, \tag{25}$$

where $\hat{\mu}_p$ and $\hat{\sigma}_p^2$ are the time-series average and variance estimate, respectively, of the investor's portfolio return $r_{p,t+1}$ over the forecast evaluation period. We can easily calculate the portfolio return $r_{p,t+1}$ for month $t+1$ as $r_{p,t+1} = w_t r_{t+1} + (1-w_t)r_{f,t+1}$, ex-post. Unlike out-of-sample R^2, the CER measure explicitly accounts for the risk taken by an investor during the out-of-sample test period. One can interpret the CER as the risk-free rate of return that an investor is willing to trade with her optimal risky portfolio.

The CER gain is the difference between the CER for the investor who uses any candidate forecast $f^{(i)}$ of the market return and the CER for an investor who uses the historical average benchmark forecast $f^{(b)}$.

Table 4 Out-of-sample R^2 (%) of Monitoring Forecasts

Forecasts	First year in evaluation sample						
	1947	1957	1967	1977	1987	1997	2007
Proposed forecast $f^{(a)}$	0.50	0.37	0.36	0.14	− 0.09	− 0.10	− 0.24
Robust monitoring machine: $f^{(m)}$	0.57	0.55	0.52	0.34	0.35	0.32	0.18
Shrinkage: $(f^{(a)}+f^{(b)})/2$	0.29	0.21	0.21	0.09	− 0.02	− 0.03	− 0.12
Traditional robust monitoring forecasts							
DMSFE (60 months, $\delta = 1.0$)	0.40	0.30	0.39	0.24	0.04	0.01	− 0.10
DMSFE (60 months, $\delta = 0.5$)	0.45	0.42	0.38	0.19	0.10	0.14	0.09
Logistic regression (feature engineering)	0.21	0.11	0.06	− 0.05	− 0.36	− 0.48	− 0.46
Logistic regression (no feature engineering)	0.37	0.23	0.25	0.15	− 0.08	− 0.30	− 0.60

This table repeats Table 1 reporting the forecasting performance of monitoring forecasts in terms of out-of-sample R^2 (%). Traditional robust monitoring forecasts are based on discounted mean square forecast error (DMSFE) from Stock and Watson (2004) and logistic regressions. Each column corresponds to a different sample split year from 1947 to 2007. All evaluation period ends in December 2017

$$(\text{CER gain of forecast } f^{(i)}) = CER^{(i)} - CER^{(b)}. \tag{26}$$

We multiply CER gain by 1,200 so it represents the annual percentage portfolio management fee that an investor would be willing to pay to have access to the given forecast $f^{(i)}$ instead of the historical average benchmark forecast $f^{(b)}$.

Table 5 repeats the analysis in Table 4, but reports the annualized CER gains; that is, the economic values of forecasting market returns by given forecasts instead of the benchmark forecast. We observe the same pattern in the outcomes, confirming the results expressed in terms of out-of-sample R^2 in Table 4. The CER gain is 1.05% from 1947 to 2017, while the combination forecast has a CER gain of 0.90%. The improvement increases when we evaluate more recent samples since 2007 (0.84 over 0.40) and since 1997 (1.22 over 0.36). Therefore, the CER gains using the robust monitoring machine are relatively stable over time.

Discussion

Out-of-sample performance

We further investigate the recent poor performance of the combination forecast by Rapach et al. (2010). We repeat the rolling-correlation tests in Rapach et al. (2010) by extending the ending date of the data from 2005 to 2017. We compute the correlations between the equity premium and 14 individual predictors in Sect. Data, benchmark, and proposed forecasts based on 10-year rolling windows. The correlation plots (Fig. 5 in Appendix 2) reveal the following. First, the correlations between realized risk premium and the 14 predictor variables that make up the combination forecasts are highly unstable since 2006. Therefore, any forecasts trained using the past predictor variable data suffer from this strong instability, which is too severe to be mitigated even by rolling-window estimation rather than expanding window. Second, we do not find that the recent poor performance is particularly linked to business cycles. We have several business cycles in the 1947–2017 sample period, but the performance in the 2007–2017 period is much worse than the rest. Therefore, the recent poor performance probably

Table 5 Certainty Equivalent Return Gains of Monitoring Forecasts

	First year in evaluation sample						
	1947	**1957**	**1967**	**1977**	**1987**	**1997**	**2007**
Proposed forecast $f^{(a)}$	0.90	0.77	0.81	0.33	0.10	0.36	0.40
Robust monitoring machine: $f^{(m)}$	1.05	1.08	1.13	0.81	0.92	1.22	0.84
Shrinkage: $(f^{(a)}+f^{(b)})/2$	0.51	0.44	0.46	0.21	0.10	0.23	0.18
Traditional robust monitoring forecasts							
DMSFE (60 months, $\delta = 1.0$)	0.69	0.59	0.76	0.45	0.26	0.30	0.54
DMSFE (60 months, $\delta = 0.5$)	0.91	0.94	0.90	0.51	0.48	0.63	0.67
Logistic regression (feature engineering)	0.54	0.35	0.44	0.30	0.06	− 0.33	− 0.61
Logistic regression (no feature engineering)	0.46	0.36	0.30	0.13	− 0.17	0.01	0.14

This table repeats Table 4 reporting the forecasting performance of monitoring forecasts, but in terms of certainty equivalent returns (%, annualized). Traditional robust monitoring forecasts are based on discounted mean square forecast error (DMSFE) from Stock and Watson (2004) and logistic regressions. Each column corresponds to a different sample split year from 1947 to 2007. All evaluation period ends in December 2017

differs from the usual business cycle narrative that return predictability is better in bad times than in good times.

We stress that we do not argue that our robust monitoring forecast is better than the combination forecast only because of their relative performance (measured by out-of-sample R^2) in the recent period (2007–2017), ignoring their previous performances. Instead, we argue that a forecast should be chosen based on not only their average performances (e.g., out-of-sample R^2 or average loss difference, $E[d]$) but also the (in)stability of their performance; for example, the variance of loss difference, $var(d)$. Suppose we live in the end of 2005 (when the test sample ends in Rapach et al. 2010) and try to choose between the robust monitoring forecast and the combination forecast. Table 6 shows that their average performances up to 2015 are practically identical, consistent with Rapach et al. (2010). However, Table 6 also shows that the combination forecast displays twice larger forecast instability than the robust monitoring forecast. The 2006–2017 period then demonstrates how dramatically a forecast with unstable performance (i.e., combination forecast) can fail in extreme times such as the financial crisis. Therefore, we consider the financial crisis as an extreme observation that tests forecast instability, not as an outlier to remove. This interesting pattern resembles the poor performance of ETFs after their inceptions (Brightman and Li 2015) and vanishing anomalies after academic publications (McLean and Pontiff 2016).

Empirical relationship to the economy

Mounting evidence shows that return predictability increases in bad economic times.[20] This established empirical fact produces hindsight theories that try to explain it and allows prediction models to take advantage of it for better ex-post performance.[21] However, the real question is if we can create a prediction model that can foresee the concentration of return predictability before the observed data

[20] See (Rapach et al. 2010; Henkel et al. 2011), and (Dangl and Halling 2012) among others.

[21] Cujean and Hasler (2017) provide theoretical model to explain why stock return predictability concentrates in bad times.

Table 6 Forecasting Performance Before and Since 2006

Forecast	1947–2005				2006–2017
	Out-of-sample R^2 (%)	$E[d.]$ ($\times 10^{-6}$)	$Var[d.]$ ($\times 10^{-8}$)	Diebold-Mariano Test (*p*-value)	Out-of-sample R^2 (%)
Robust (m)	0.65	11.2	0.275	0.004	0.17
Proposed (a)	0.64	11.1	0.519	0.040	− 0.25

This table reports forecasting performance measures for two sub-periods: 1947–2005 and 2006–2017. $E[d_i]$ and $Var[d_i]$ for $i \in \{a, m\}$ denote the expected value and variance of loss difference, which represent average forecasting performance and its relative risk, respectively. Diebold-Mariano tests are to show if a forecast has significantly different predictability from a benchmark. All forecasts are free from look-ahead bias

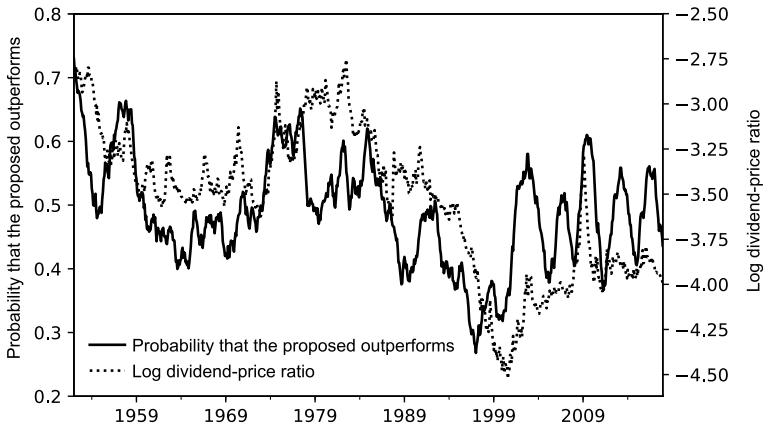

Fig. 3 Return Predictability and Stock Valuations This figure shows the log dividend-price ratio (dotted line) and the probability that the proposed forecast outperforms the benchmark, computed by robust monitoring machine (solid line). The probability in the figure is shown as its 24-month trailing moving-average. The horizontal axis is year

reveals it. To answer this question, we compare our robust monitoring machine outputs with two variables that represent bad economic times.

We first consider a variable for stock valuation. The log dividend-price ratio of S &P500 is commonly used in prior studies as the valuation ratio and a return predictor. High dividend-price ratios mean low valuation of stocks and thus bad economic times. Figure 3 displays both the log divided-price ratio (dotted line) and the computed real-time probability p_t (solid line) that the proposed forecast $f^{(a)}$ will outperform the benchmark $f^{(b)}$, calculated by the robust monitoring machine. The fluctuation patterns of two graphs are similar. When stock valuation is low, the robust monitoring machine favors the proposed forecast over the benchmark. However, determining whether the dividend-price ratio is low or high is often a hindsight observation because the true long-run mean of the dividend-price ratio is unknown, For example, the divided-price ratio is systematically lower in the last two decades than in the twentieth century because of either a permanent regime change or temporal abnormality. By contrast, the computed real-time probability p_t is always measured between zero and one so that a forecaster can easily interpret its meaning and magnitude.

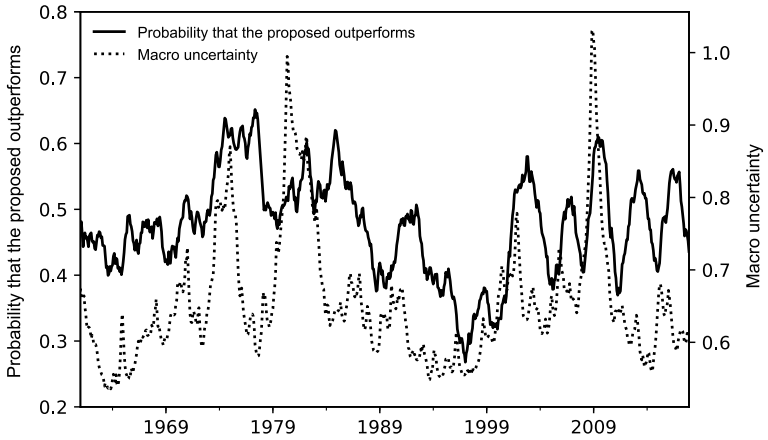

Fig. 4 Return Predictability and Macroeconomic Uncertainty This figure shows the macroeconomic uncertainty (dotted line) and the probability that the proposed forecast outperforms the benchmark, computed by robust monitoring machine (solid line). The probability in the figure is shown as its 24-month trailing moving-average. The horizontal axis is year

Similarly, Fig. 4 plots the macro uncertainty measure from Jurado et al. (2015) (dotted line) and the probability p_t (solid line). When macro uncertainty is high (usually during bad economic times), the robust monitoring machine favors the proposed forecast over the benchmark. This macro uncertainty measure consists of the principal components that require a full-sample estimation. By contrast, the computed real-time probability p_t captures the same information in real time. From these two plots, we confirm that our robust monitoring machine can foresee the concentration of return predictability before the observed data explicitly reveals it to econometricians, although high correlations between p_t and market cycles are not surprising per-se.

Factors in the cross-sectional variation of stock returns

Starting from the market factor, researchers proposed hundreds of new factors (or anomalies) for cross-sectional variation of stock returns, such as information discreteness by Da et al. (2014) or betting-against-beta by Novy-Marx and Velikov (2022), among many others.[22]

We find that not all factors are free from look-ahead bias. Some factors are naturally subject to look-ahead bias because of their empirical construction. Some factors are impossible to construct in real-time because of poor database availability in early days. Finally, some factors have insufficient statistical evidence in early periods; therefore, a portfolio strategy based on such factors may look unwise to investors in those periods. The last type of bias is well explained by Martin and Nagel (2022).

Although look-ahead bias does not directly affect formal asset pricing tests, the performance of portfolio strategies derived from the proposed anomalies and factors can mislead readers in the presence of look-ahead bias. The framework proposed in this paper can potentially help researchers distinguish factors with or without look-ahead bias. However, this task is crucially important and sensitive, so it is beyond the scope of this study, and we leave it for future research.

[22] See (Harvey et al. 2016; McLean and Pontiff 2016), and (Hou et al. 2020) for a complete list of factors and anomalies studied in the literature.

Trustworthy machine learning solutions

The underlying topic in this study extends to general requirements for trustworthy machine learning. We emphasize the pursuit of robustness and eliminating look-ahead biases. Such desirable qualities are pre-requisites for trustworthy machine learning. Similarly, Holzinger (2021) argues that trustworthy machine learning solutions, or Artificial Intelligence (AI) solutions in a broader sense, should bear other qualities such as comprehensibility, explainability, and interpretability for the human expert, in addition to robustness. Holzinger (2021) emphasizes the importance of human experts in decision processes using artificial intelligence system to achieve such qualities. We relate our study to the lesson from Holzinger (2021), as our approach provides a way of curbing unintentional biases by human experts.

Conclusion

Forecasting is a cornerstone for decision-making. Recently, the superior forecasting performance of machine learning methods drew significant attention. However, the black-box nature of machine learning methods often exacerbates the out-of-sample R^2-hacking problem, which exaggerates the true forecasting performance through over-fitting. In contrast, this study exploits the black-box nature of machine learning methods to avoid out-of-sample R^2-hacking.

We provide a machine-learning solution for the out-of-sample R^2-hacking problem in robust monitoring. The resulting forecast improves the average performance of a proposed forecast by 15% and reduces its variance of performance by 46.3%. The DM test statistic becomes significant, with its p-value falling from 0.068 to 0.004. The sensitivity and specificity of monitoring as a classifier are 57.0% and 56.8%, respectively, and are statistically different from those of random classifiers. The robust monitoring machine predicts the time-variation of return-predictability over business cycles without look-ahead bias.

The proposed forecast in our application is a combination forecast from Rapach et al. (2010), yet we can apply the robust monitoring machine to any forecast to improve its performance and robustness. Therefore, professional forecasters can use our approach as a final touch to any sophisticated prediction model they choose. Our approach facilitates other forecasting methods instead of competing with them. Additionally, our three-step approach can be implemented in many different ways for further practical improvements.

The framework of the robust monitoring machine can be applied to other types of forecasting examples, such as predicting macro-economic variables or corporate earnings. Furthermore, the underlying idea of our approach can extend to other fields in finance. For example, we can construct a trading strategy for a mutual fund manager whose performance is evaluated based on the benchmark portfolio. Using machine learning techniques, a fund manager can deviate from the benchmark only when the signal is strong enough to earn extra trading profits above a pre-determined threshold. We leave such extensions for future research.

Appendix 1: A list of individual predictor variables

We list the names and brief descriptions of 14 predictor variables we use as follows:

1. Dividend-price ratio (dp): log of a 12-month moving sum of dividends paid on the S &P 500 Index minus the log of stock prices (S &P 500 Index)
2. Dividend yield (dy): log of a 12-month moving sum of dividends minus the log of lagged stock prices.
3. Earning-price ratio (ep): log of a 12-month moving sum of earnings on the S &P 500 Index minus the log of stock prices.
4. Dividend-payout ratio (de): log of a 12-month moving sum of dividends minus the log of a 12-month moving sum of earnings.
5. Equity risk premium volatility (rvol): based on a 12-month moving standard deviation estimator
6. Book-to-market ratio (bm): book-to-market value ratio for the Dow Jones Industrial Average
7. Net equity expansion (ntis): ratio of a 12-month moving sum of net equity issues by NYSE-listed stocks to the total end-of-year market capitalization of New York Stock Exchange (NYSE) stocks.
8. Treasury bill rate (tbl): interest rate on a three-month Treasury bill (secondary market).
9. Long-term yield (lty): long-term government bond yield.
10. Long-term return (ltr): return on long-term government bonds.
11. Term spread (tms): long-term yield minus the Treasury bill rate.
12. Default yield spread (dfy): difference between Moody's BAA- and AAA-rated corporate bond yields.
13. Default return spread (dfr): long-term corporate bond return minus the long-term government bond return.
14. Inflation (infl): calculated from the CPI for all urban consumers.

Appendix 2: Time-varying correlation of predictor variables

See Fig. 5

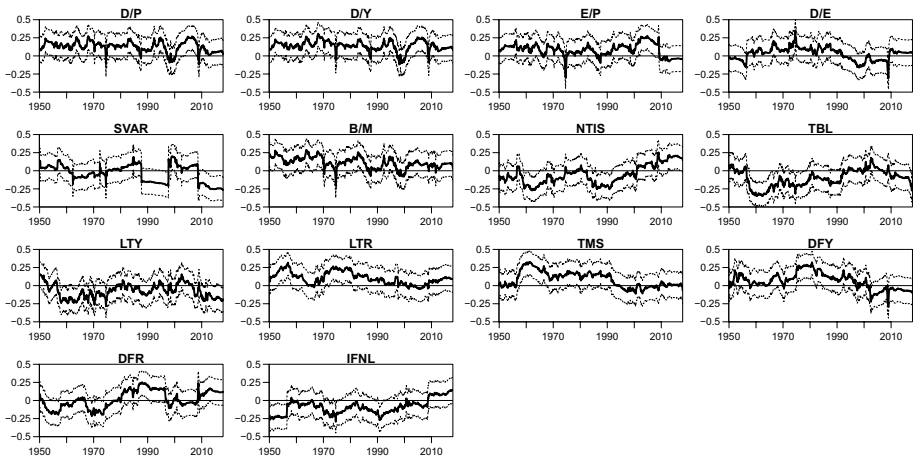

Fig. 5 Correlations between the equity premium and individual predictors based on 10-year rolling windows The figure repeats the rolling-correlation tests of Rapach et al. (2010) by extending the ending date of the data from 2005 to 2017. The date on the horizontal axis gives the end date of the 10-year period. Dashed lines indicate 95% confidence intervals

Appendix 3: Machine learning algorithms used in the paper

1. *Random Forest* This is an bagging method that combines forecasts from many small decision trees. This algorithm introduces extra randomness when growing trees. Instead of trying to fit the whole sample, each small decision tree will only use a sub-sample of the dataset. Averaging the results from small decision trees can improve the predictive accuracy and control over-fitting. This brings greater diversity, which trades a higher bias for a lower variance and yields an overall better model. This algorithm repeats the process for multiple times while producing one decision tree for each time. At the end, it combines multiple decision trees and averages their forecasts. We use default loss function Gini impurity of RandomForest in scikit-learn package.
2. *2. Extremely Randomized Trees* Instead of using sub-sample of the dataset and searching for the best possible threshold for each feature when splitting a node, this algorithm uses the full sample and random thresholds for each feature. This trades more biaes for a lower variance and usually will be faster to train than regular random forecast since finding the best possible threshold for each feature at every node is very time-consuming. We use default loss function Gini impurity of ExtraTrees in scikit-learn package.
3. *3. Gradient Boosting* This algorithm works by sequentially adding a new decesion tree to an ensemble of previous trees with each new one trying to correct the forecasting errors from its predecessor. It fit the new predictor to the residual errors made by the previous predictor. Shallow trees on their own are "weak learners" with weak predictive power. The theory behind boosting suggests that many weak learners may, as an ensemble, comprise a single "strong learner" with greater stability than a single complex tree. We use default loss function friedman-mse of GradientBoosting in scikit-learn package.

Abbreviations

TPR	True positive rate
TNR	True negative rate
PPV	Positive predicted value
NPV	Negative predictive value
ACC	Accuracy
CER	Certainty equivalent return
AI	Artificial intelligence

Acknowledgements
The authors thank Hitesh Doshi, Cao Fang (2021 Asian FA discussant), Kris Jacobs, Jun Myung Song, Raul Susmel, and seminar participants at 2021 Asian FA Annual Meeting, 2021 AFAANZ Conference, and University of Houston for their valuable feedback.

Author contributions
All authors contributed equally to this work. All authors have read and approved the final manuscript.

Funding
No funding was received for conducting this study.

Availability of data and materials
The datasets analyzed in the current study are available at Amit Goyal's website: http://www.hec.unil.ch/agoyal

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 26 June 2022 Accepted: 18 April 2023
Published online: 11 July 2023

References

Abad-Segura E, González-Zamar M-D (2020) Global research trends in financial transactions. Mathematics 8:614
Abel AB (1990) Asset prices under habit formation and catching up with the joneses. Am Econ Rev 80:38
Bailey DH, Ger S, de Prado ML, Sim A (2015) Statistical overfitting and backtest performance. In: Risk-based and factor investing. Elsevier, pp 449–461
Bates JM, Granger CWJ (1969) The combination of forecasts. J Oper Res Soc 20:451–468
Brightman C, Li F, Xi L (2015) Chasing performance with ETFs, Research Affiliates Fundamentals (November)
Campbell JY, Thompson SB (2008) Predicting excess stock returns out of sample: Can anything beat the historical average? Rev Financ Stud 21:1509–1531
Chordia T, Goyal A, Saretto A (2017) p-hacking: Evidence from two million trading strategies
Christ M, Braun N, Neuffer J, Kempa-Liehr AW (2018) Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh - A Python package). Neurocomputing 307:72–77
Cujean J, Hasler M (2017) Why does return predictability concentrate in bad times? J Finance 72:2717–2758
Da Z, Gurun UG, Warachka M (2014) Frog in the pan: continuous information and momentum. Rev Financ Stud 27:2171–2218
Dangl T, Halling M (2012) Predictive regressions with time-varying coefficients. J Financ Econ 106:157–181
Prado D, López M (2018) The 10 reasons most machine learning funds fail. J Portfolio Manag 44:120–133
de Prado ML (2019) A data science solution to the multiple-testing crisis in financial research. J Financ Data Sci 1:99–110
Diebold FX, Mariano RS (2002) Comparing predictive accuracy. J Bus Econ Stat 20:134–144
Feng G, Giglio S, Xiu D (2020) Taming the factor zoo: a test of new factors. J Financ 75:1327–1370
Ferreira MA, Santa-Clara P (2011) Forecasting stock market returns: the sum of the parts is more than the whole. J Financ Econ 100:514–537
Freyberger J, Neuhierl A, Weber M (2020) Dissecting characteristics nonparametrically. Rev Financ Stud 33:2326–2377
Gali J (1994) Keeping up with the joneses: consumption externalities, portfolio choice, and asset prices. J Money Credit Bank 26:1–8
Gibbs C, Vasnev AL (2018) Conditionally optimal weights and forward-looking approaches to combining forecasts, Available at SSRN 2919117
Goldstein I, Spatt CS, Ye M (2021) Big data in finance. Rev Financ Stud 34:3213–3225
Goyal A, Welch I (2008) A comprehensive look at the empirical performance of equity premium prediction. Rev Financ Stud 21:1455–1508
Granziera E, Sekhposyan T (2019) Predicting relative forecasting performance: an empirical investigation. Int J Forecast 35:1636–1657
Gu S, Kelly B, Xiu D (2020) Empirical asset pricing via machine learning. Rev Financ Stud 33:2223–2273
Harvey CR, Liu Y, Zhu H (2016) and the cross-section of expected returns. Rev Financ Stud 29:5–68
Heaton JB, Polson NG, Witte JH (2017) Deep learning for finance: deep portfolios. Appl Stoch Model Bus Ind 33:3–12
Hendry DF, Clements MP (2004) Pooling of forecasts. Econom J 7:1–31
Henkel SJ, Spencer Martin J, Nardari F (2011) Time-varying short-horizon predictability. J Financ Econ 99:560–580
Holzinger A (2021) The next frontier: ai we can really trust, in machine learning and principles and practice of knowledge discovery in databases - international workshops of ECML PKDD 2021. Springer, pp 427–440
Hou K, Xue C, Zhang L (2020) Replicating anomalies. Rev Financ Stud 33:2019–2133
Inoue A, Kilian L (2005) In-sample or out-of-sample tests of predictability: Which one should we use? Econom Rev 23:371–402
Inoue A, Kilian L (2006) On the selection of forecasting models. J Econom 130:273–306
Jurado K, Ludvigson S, Ng S (2015) Measuring uncertainty. Am Econ Rev 105:1177–1216
Klibanoff P, Marinacci M, Mukerji S (2005) A smooth model of decision making under ambiguity. Econometrica 73:1849–1892
Kou G, Chao X, Peng Y, Alsaadi FE, Herrera-Viedma E (2019) Machine learning methods for systemic risk analysis in financial sectors. Technol Econ Dev Econ 25:716–742
Kou G, Peng Y, Wang G (2014) Evaluation of clustering algorithms for financial risk analysis using mcdm methods. Inf Sci 275:1–12
Kou G, Yong X, Peng Y, Shen F, Chen Y, Chang K, Kou S (2021) Bankruptcy prediction for smes using transactional data and two-stage multiobjective feature selection. Decis Support Syst 140:113429
Kou G, Yüksel S, Dinçer H (2022) Inventive problem-solving map of innovative carbon emission strategies for solar energy-based transportation investment projects. Appl Energy 311:118680
Li T, Kou G, Peng Y, Yu Philip S (2021) An integrated cluster detection, optimization, and interpretation approach for financial data. IEEE Trans Cybern 52:13848–13861
Martin IWR, Nagel S (2022) Market efficiency in the age of big data. J Financ Econ 145:154–177
McLean RD, Pontiff J (2016) Does academic research destroy stock return predictability? J Financ 71:5–32
Novy-Marx R, Velikov M (2022) Betting against betting against beta. J Financ Econ 143:80–106
Pesaran MH, Timmermann A (1995) Predictability of stock returns: robustness and economic significance. J Financ 50:1201–1228
Rapach DE, Strauss JK, Zhou G (2010) Out-of-sample equity premium prediction: combination forecasts and links to the real economy. Rev Financ Stud 23:821–862
Rapach DE, Strauss JK, Zhou G (2013) International stock return predictability: What is the role of the united states? J Finance 68:1633–1662

Roll R (1992) A mean/variance analysis of tracking error. J Portfolio Manag 18:13–22
Stock JH, Watson MW (2004) Combination forecasts of output growth in a seven-country data set. J Forecast 23:405–430
Timmermann A (2006) Forecast combinations. Handb Econ Forecast 1:135–196
Yae J (2018) The efficient frontier of forecasts: Beyond the bias-variance tradeoff, Working Paper
Yae J (Forthcoming) Unintended look-ahead bias in out-of-sample forecasting, Applied Economics Letters
Zhu X, Zhu J (2013) Predicting stock returns: a regime-switching combination approach and economic links. J Bank Finance 37:4120–4133
Zhu Yi, Timmermann A (2017) Monitoring forecasting performance, UCSD working paper

Publisher's Note