Journal of Commodity Markets xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Commodity Markets

journal homepage: www.elsevier.com/locate/jcomm

Gold risk premium estimation with machine learning methods☆

Juan D. Díaz [a], Erwin Hansen [b,*], Gabriel Cabrera [c]

[a] Department of Management Control and Information Systems, Faculty of Economics and Business, University of Chile, Diagonal Paraguay 257, Of., 2001, Santiago, Chile
[b] Department of Business Administration, Faculty of Economics and Business, University of Chile, Diagonal Paraguay 257, Of., 1204, Santiago, Chile
[c] Faculty of Economics and Business, University of Chile, Diagonal Paraguay 257, Santiago, Chile

ARTICLE INFO

Keywords:
Gold
Risk premia
Machine learning
Big data
Combination forecasts
Portfolios

ABSTRACT

This paper assesses the accuracy of several machine learning models' predictions of the gold risk premium when using an extensive set of 186 predictors. We perform an out-of-sample evaluation and consider both statistical and portfolio metrics. Our results show that machine learning methods and forecast combinations have a limited ability to outperform the historical mean when predicting the gold risk premium. Slightly better results are obtained when predictors are used individually. More specifically, we find that several technical indicators (moving average and momentum series) have forecasting power during periods of expansion, while several business cycle variables and geopolitical risk variables help predict the gold risk premium during recessions. An economic evaluation accounting for transaction costs shows that investors using machine learning methods to estimate expected returns on gold should anticipate limited portfolio gains.

1. Introduction

There has been increasing scholarly interest in the use of gold as an investment asset, with a particular focus on its diversification, hedging, and safe heaven features (see, e.g., Baur et al., 2020; and O'Connor et al., 2015). However, much less emphasis has been placed on the gold risk premium and its drivers. Exceptions in this regard are studies by Nguyen et al. (2019) and Dichtl (2020). The former estimates a parsimonious gold risk premium model and identifies both the variance risk premium and jump risk as its main drivers, while the latter documents the forecasting power of variables such as the default spread.

Moreover, there is growing interest in using a combination of machine learning (ML) methods and big data, i.e., a large set of predictors, in finance applications. For example, Huck (2019) and Krauss et al. (2017) examine the performance of several ML methods when these are implemented to evaluate statistical arbitrage strategies using a large set of predictors obtained from U.S. stock returns. Their results show that it may be beneficial for investors to leverage these kinds of models when planning long-short portfolio strategies. In the smaller body of literature specifically examining risk premium estimates, there is also a swelling tide of interest in ML applications. For instance, Gu et al. (2020) conducted an extensive empirical study of the equity risk premium in the U.S. that compared several ML methods, Bianchi et al. (2020) researched bond risk premia, and Wu et al. (2020) investigated hedge fund risk

☆ We thank comments from participants at the Hamburg Business School's finance seminar, World Finance Conference 2021, and 4th International Conference on Econometrics and Statistics (EcoSta 2021).

* Corresponding author.
E-mail addresses: juadiaz@fen.uchile.cl (J.D. Díaz), ehansen@fen.uchile.cl (E. Hansen), gcabrerag@fen.uchile.cl (G. Cabrera).

https://doi.org/10.1016/j.jcomm.2022.100293
Received 21 April 2022; Received in revised form 30 September 2022; Accepted 11 October 2022
Available online 15 October 2022

premia using ML methods.[1]

This article aims to build on the existing literature by evaluating the gold risk premium forecasting performance of several ML methods and a large set of predictors. Specifically, we evaluate the following ML algorithms in our empirical exercise: (i) regularization methods for linear regression such as LASSO, Ridge, and Elastic Net; (ii) decision learning methods such as random forests (RF) and gradient boosted regression trees (GBRT); and (iii) neural networks (NN).[2] Furthermore, since previous literature has shown that forecast combinations usually outperform individual forecasting models (see, e.g., Rapach et al., 2010), we also consider several forecast combinations. We measure the gold risk premium as the difference between continuously compounded gold returns over a risk-free rate. Gold returns are computed using end-of-month spot gold fixing prices from the London Bullion Market (3:00 PM, London time) in USD, and the risk-free rate is the treasury bill rate provided by Goyal and Welch (2008). As mentioned above, our empirical exercise is a big data exercise in which we consider a set of 186 predictors, including financial and macroeconomic variables, uncertainty variables, as well as technical indicators relating to stock index and gold returns. To our knowledge, this is the largest set of variables evaluated in the relevant literature so far. Most of the predictor variables are economically motivated in the stock, bond, and commodity return predictability literature (see, e.g., Çakmaklı and Van Dijk, 2016; Bianchi et al., 2020; Gargano and Timmermann, 2014) as significant predictors of financial asset risk premia. Neely et al. (2014) and Baur et al. (2020) also provide evidence that technical indicators can be used as predictors of risk premia.

We perform both statistical and economic assessments of the gold risk premium forecasting performance of these ML methods (when these are used in conjunction with a large set of predictors). The statistical evaluation is based on the out-of-sample R^2 metric introduced by Goyal and Welch (2008) in their influential study on stock return predictability. We use the forecast accuracy test of Clark and West (2007) to provide formal comparisons between the potential gains from each model's predictions and those obtained from a benchmark model. The economic evaluation is motivated by prior evidence showing that a model may deliver significant economic gains despite offering a poor statistical fit (see, e.g., Kandel and Stambaugh, 1996; and Cenesizoglu and Timmermann, 2012). We conduct the economic evaluation of the ML methods using portfolios built from three assets: gold, the S&P 500 index, and a risk-free asset. These portfolios use as inputs gold risk premium estimates from each of the models under evaluation. The portfolios are estimated recursively, month by month, and the results are used to compute several performance metrics: the Sharpe ratio (SR), certainty equivalent returns (CER), and performance fee (Δ) estimates accounting for transaction costs (see e.g., Çakmaklı and Van Dijk, 2016), among others.

Our results show that, out-of-sample (OOS) and based on the set of predictors we considered, the ML methods were unable to outperform the benchmark model, i.e., the historical mean of gold excess returns. Neither the ML methods individually, nor the forecast combinations delivered better predictions than the benchmark model's. There was one exception nonetheless: the partial least squares regression (PLSR) method offered slightly better forecasting accuracy. Overall, these results are in line with those reported by Goyal and Welch (2008) for stock index returns and they also cohere with those obtained by Dichtl (2020) for gold returns using linear models. As such, the simultaneous use of a large set of predictors does not seem to contribute to a significant improvement in the forecasting performance of the models under evaluation. Our findings are also consistent with those reported by Hollstein et al. (2021). These authors focused on several commodities and concluded that gold prices are difficult to predict out-of-sample. They used a smaller set of macroeconomic variables than ours though, as well as forecast combinations. Our results are also similar to those obtained by Huck (2019), who investigated stock return predictability using ML methods. Huck concluded that the use of a larger set of predictors damages predictive performance. From an economic perspective, several ML methods (e.g., LASSO, Elastic Net, and Random Forest) can deliver superior portfolio performance than the benchmark model according to the SR and CER metrics. However, when transaction costs are taken into account, these portfolio gains become negligible. Baur et al. (2020) report a similar result based on the evaluation of more than 4000 gold trading strategies.

In a second empirical exercise, we evaluate the forecasting performance of individual variables and obtain better results than when using the ML methods. Indeed, we identify several variables that are significant predictors of the gold risk premium. We find that aggregate variables, such as the default spread, housing starts, the unemployment rate, and the producer price index (PPI), as well as some technical indicators, can help predict future gold returns. Oil prices and geopolitical risk variables also are informative for estimating gold risk premia. Furthermore, we document an interesting pattern when examining periods of expansion and recession separately. We observe that macroeconomic variables have the most forecasting power during recessions, but that technical indicators provide the most predictive power during periods of expansion. Most of these predictors have also been identified in prior studies. Hence, it is unlikely that our results were obtained by chance, i.e., simply due to the sheer number of predictors (see, e.g., Dichtl, 2020; Hollstein et al., 2021; Aye et al., 2015; Pierdzioch et al., 2014; Le and Chang, 2012; Tanin et al., 2022; Baur and Smales, 2020; Gozgor et al., 2019). When evaluating these individual predictors economically, we obtain mixed results. According to the CER and SR estimates, several of them outperform the benchmark model. Yet, the opposite is true when referring to the portfolio risk and turnover metrics instead. When considering performance fees net of transaction costs, we find that several gold return technical indicators still deliver portfolio gains.

We aim to contribute to the literature investigating the advisability of combining ML methods and big data (i.e., a large set of predictors) in finance applications. We add to research by Gu et al. (2020), Bianchi et al. (2020), and Wu et al. (2020), who uses ML methods to predict stock, bond, and hedge fund risk premia, by focusing instead on a relatively understudied financial asset: gold. More generally, we also seek to build on research by Huck (2019) and Krauss et al. (2017), among others, since these authors have used ML

[1] See Weigand (2019) for a recent survey of this literature.

[2] See Gambella et al. (2021) for a detailed survey of ML methods.

methods to evaluate the performance of several financial arbitrage strategies.

Our study is also connected to the literature on gold price and return forecasting (see, e.g., Pierdzioch et al., 2014, 2015, 2016a, 2016b; and Aye et al., 2015). Nevertheless, our paper differs significantly from this scholarship due to our choice of forecasting methods and our set of predictors. The study conducted by Malliaris and Malliaris (2016) is more closely related to this one because it forecast gold returns using a decision tree model. Nevertheless, we consider more sophisticated decision tree methods (random forest and GBRT models) and perform an out-of-sample evaluation rather than an in-sample analysis as they did. Jabeur et al. (2021) also used gradient boosting methods to forecast gold prices but from an in-sample perspective only. Two further related studies are those by Pierdzioch and Risse (2020) and Risse (2019). In the former, the authors used a multivariate regression tree to forecast the returns on four precious metals, including gold. Their multivariate forecasts outperformed their univariate ones. In the latter, Risse (2019) forecast gold returns by combining wavelet decomposition and an ML algorithm (support vector regression). However, this article considers several additional ML methods and tests a broader set of predictors.

The remainder of this paper is structured as follows. In section 2, we describe our dataset; in section 3, we describe the ML algorithms and the forecasting framework; we report our empirical results in section 4; we perform several robustness tests in section 5; and finally, we conclude in section 6.

2. Data

Our dataset is monthly and covers the period from January 1970 until December 2019. We measure the gold risk premium as the difference between continuously compounded gold returns over a risk-free rate. Gold returns are computed using end-of-month spot gold fixing prices from the London Bullion Market (3:00 p.m., London time) in USD,[3] and the risk-free rate is the treasury bill rate provided by Goyal and Welch (2008).[4] Fig. 1 plots gold excess returns (upper panel) and gold cumulative excess returns (lower panel). For reference purposes, the lower panel also reports the S&P 500's cumulative excess returns over the same period. Gray bars indicate recession periods, as defined by the National Bureau of Economic Research (NBER). While the S&P 500 cumulative return has trended upwards since the 1980s, gold cumulative returns have followed a mean-reverting pattern.[5]

We compiled data from several sources to produce a set of 186 possible drivers of the gold risk premium. To the best of our knowledge, this is the largest set of predictors in the relevant literature so far (see, e.g., Nguyen et al., 2019; Baur et al., 2020; Dichtl, 2020, Pierdzioch et al., 2014 and 2015). This large number of potential predictors warrants the use of machine learning techniques in our empirical analysis below. The set of 186 predictors includes those listed by Christiansen et al. (2012), which are sorted into the following categories: equity market and risk factor variables; interest rates, spreads, and bond market factors; foreign exchange variables and risk factors; liquidity and credit risk variables; and macroeconomic variables. We managed to update 33 of the 38 original predictors.[6] To this dataset, we add three additional equity factor series retrieved from Professor Kenneth French's online data library: the momentum (MOM), profitability (RMW), and investment (CMA) factors. Furthermore, we add 112 additional macroeconomic variables from the database by McCracken and Ng (2016). This large macroeconomic database is used by Bianchi et al. (2020) in their study of bond risk predictability using machine learning. In addition, we include 12 variables capturing macroeconomic and financial uncertainty: the option-implied volatility VIX index and VXO index, the risk aversion and financial uncertainty indices of Bekaert et al. (2020), the geopolitical risk indices (3) of Caldara and Iacoviello (2018), the economic policy uncertainty (EPU) index of Baker et al. (2016), the financial uncertainty indices (3) of Jurado et al. (2015) and the monetary policy uncertainty index of Husted et al. (2020). Economic uncertainty, as measured by all these indices, has been shown to impact macroeconomic aggregates and financial markets. For example, Gkillas et al. (2020), Baur and Smales (2020), and Triki and Maatoug (2020) document a significant relationship between geopolitical risk and gold prices.

Neely et al. (2014) show that technical indicators provide more valuable information than macroeconomic variables when forecasting the U.S. equity risk premium and that this information is both statistically and economically significant. What is more, the authors demonstrate that both are complementary since they are sensitive to different stages of the business cycle. Baur et al. (2020) study the ability of technical indicators to identify investment signals in the gold market. The authors show that market timing based on these indicators does not consistently produce significant economic gains. Motivated by this evidence, we include 19 technical indicators relating to stock returns and 13 gold technical indicators (six moving average series, two momentum series, and five realized volatility series; for the stock returns, we also include six volume series) as possible drivers of gold risk premia. Each technical indicator defines a signal (S) that takes a value of 1 or 0 for long and short positions, respectively. The specific definitions of this set of signals are as follows. A moving average (MA) signal is given by:

[3] We obtained this series from the Federal Reserve Economic Data (FRED) database.

[4] This gold spot price has been used by Baur et al. (2016), Baur et al. (2020), Dichtl (2020), Pierdzioch et al. (2014, 2015, 2016a, 2016b), and Pierdzioch and Risse (2020), among others. Therefore, our results may be compared to these authors'. A related paper that uses futures prices instead of spot prices to estimate the gold risk premium is that by Nguyen et al. (2019). As mentioned above, the main advantage of using spot prices rather than futures prices is the potential for comparisons with prior literature. An additional advantage is that spot prices are a recognized benchmark in the gold industry (see, e.g., World Gold Council). Nevertheless, a disadvantage is that transactions are mostly completed on OTC markets, where price transparency is lower than on futures markets (see Batten and Lucey, 2010).

[5] Erb and Harvey (2017) and Erb et al. (2020) provide a detailed discussion of the time variation of gold prices based on the concept of "the gold constant", i.e., the idea that the purchasing power of gold remains relatively constant, and that inflation is its main driver.

[6] We do not include several FX factors (AFD, DOL, CT, FX bid-ask spread) and S&P500 Turnover data.

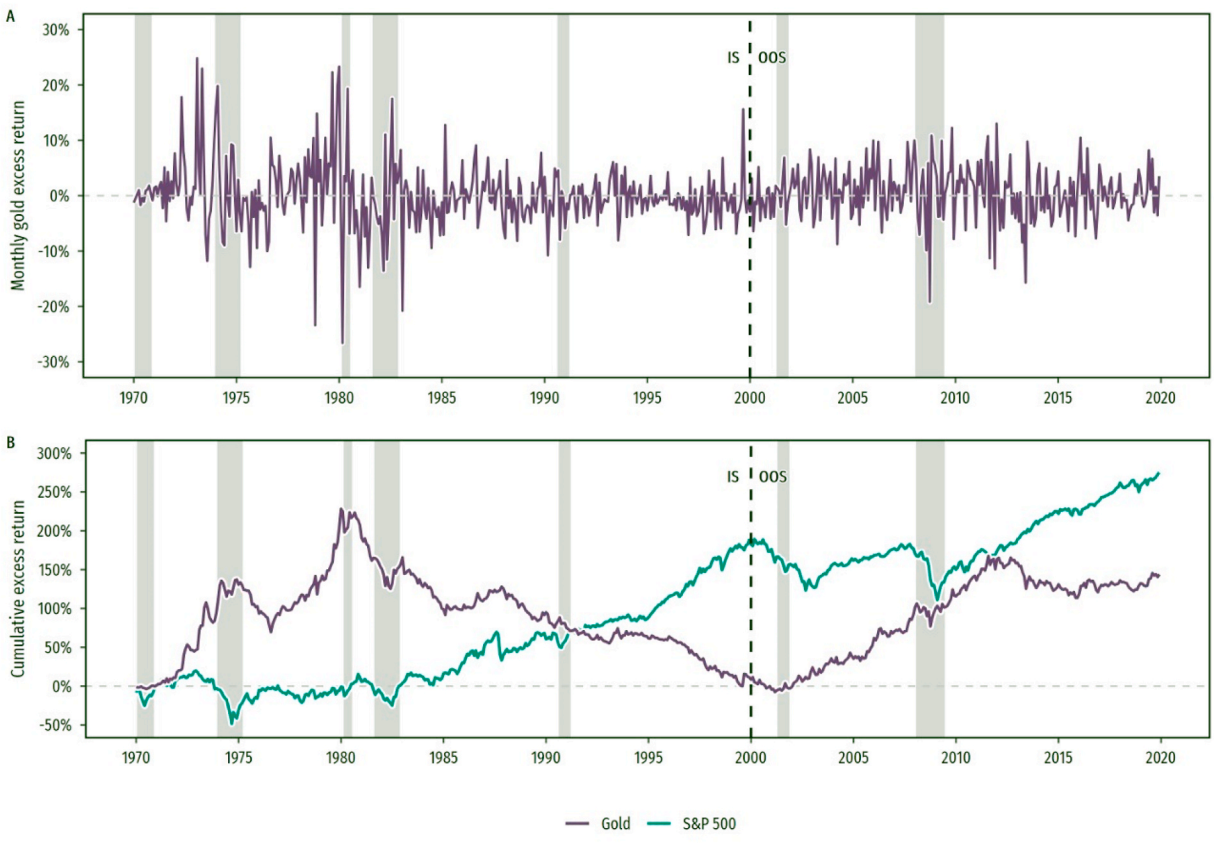

Fig. 1. Time series of Gold Returns. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

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

where

$$MA_{j,t} = \left(\frac{1}{j}\right) \sum_{i=0}^{j-1} P_{t-1}$$

for $j = s,l$; where P_t is the level of asset price, and $s(l)$ is the length of the short (long) MA ($s < l$). We considered MA rules with $s = 1,2,3$ and $l = 9,12$.

The momentum strategy $MOM(m)$ defines the following signal:

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

2.1. We evaluated $m = 9,12$

We incorporate technical indicators based on the volume of stock returns. Let OBV_t be the "on-balance" volume and defined as:

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

Where VOL_k is the trading volume during period k, and D_k is a dummy variable that takes a value of 1 if $P_k - P_{k-1} \geq 0$ and -1 otherwise. The trading signal is then given by:

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

where

$$MA_{j,t}^{OBV} = \left(\frac{1}{j}\right)\sum_{i=0}^{j-1} OBV_{t-i}$$

for $j = s, l$. We compute signals for $s = 1, 2, 3$ and $l = 9, 12$, denoted by $VOL(s, l)$.

Finally, we define a volatility clustering strategy as follows:

$$S_{i,t} = \begin{cases} 1 \text{ if } RV_t \geq RV_{t-k} \\ 0 \text{ if } RV_t < RV_{t-k} \end{cases}$$

where RV_t is a measure of monthly realized volatility during month t. We compute signals for $k = 1, 3, 6, 9$ and 12 months. In the appendix, we provide a detailed list of all the variables and their descriptive statistics.

3. Machine learning methods and empirical setting

This section presents the machine learning methods employed to forecast the gold risk premium using a large set of predictors. To assess the accuracy of these methods' predictions, we use the historical mean of the gold risk premium as a forecasting benchmark. Importantly, we also evaluate the accuracy of the forecast combinations (incorporating both the ML method forecasts and the benchmark predictions) following the combination approach developed by Rapach et al. (2010).

3.1. Machine learning methods

3.1.1. Regularization methods

We consider three types of penalized linear regression: Ridge (Hoerl and Kennard, 1970); Lasso (Tibshirani, 1996); and Elastic Net (Zou and Hastie, 2005). These methods build on the classic linear regression, but the coefficient estimates are obtained by minimizing a penalized residual sum of squares. Specifically, consider the following model:

$$R = X\beta + \varepsilon \tag{1}$$

where R is a vector that includes the outcome variable of interest (gold risk premium), X is a matrix comprising the predictors included in the model, β is a vector of coefficients to be estimated, and ε is the vector of error terms, each one independent and identically distributed (i.i.d.), with a mean of zero and constant variance. The Elastic Net method estimates β by solving the following optimization problem:

$$\min_{\beta} \|R - X\beta\|_2^2 + \lambda \bullet \left(\alpha \bullet \|\beta\|_1 + (1-\alpha) \bullet \|\beta\|_2^2\right) \tag{2}$$

for $\lambda > 0$ and $0 \leq \alpha \leq 1$, where $\| \bullet \|_1$ is the l^1 norm and where $\| \bullet \|_2$ is the Euclidean norm. In this setting, the tuning parameter λ is selected via a time series cross-validation approach and α is set to 0.5, as in Gu et al. (2020) and Drobetz et al. (2021).

It is important to note that if we set $\alpha = 0$, then the optimization problem in (2) corresponds to that of a Ridge regression, while if we set $\alpha = 1$, it corresponds to that of a Lasso. Regardless of whether $\alpha = 0$ or $\alpha = 1$, the tuning parameter λ is selected via a time series cross-validation approach.

3.1.2. Tree-based methods

We consider two machine learning methods based on trees: the Random Forest (RF, Ho, 1995; Amit and Geman, 1997; Breiman, 2001) and Gradient Boosting Regression Tree (GBRT, Hastie et al., 2009, chapter 10).

Both methods are ensemble statistical methods that extend the single regression tree (Breiman et al., 1984) by combining predictions from multiple regression trees. Specifically, RFs exploit the idea of bootstrap aggregation (bagging, Breiman, 1996) by computing their final prediction as the average of those from several single regression trees. These single regression trees are adjusted with random subsets of training data and predictors. Meanwhile, based on the idea of boosting (Freund and Schapire, 1996; Schapire, 2003), GBRTs provide better prediction accuracy than single regression trees by sequentially decreasing the bias of simple regression trees with higher bias but low variance, where the contemporary ensemble regression tree model internalizes the knowledge acquired from previously fitted models.

In practice, tuning parameters must be set for these methods to be implemented. For the RF, we set the randomly selected predictors at each split via a time series cross-validation approach; we aggregate 1000 regression trees; and we estimate trees with five terminal nodes (their complexity parameter). For the GBRT, we assemble 1000 regression trees; we use a learning rate of 0.01; and we set the complexity of the regression trees to be assembled via a time series cross-validation approach.

3.1.3. Neural networks

Our analysis also considers artificial neural networks to allow for a degree of nonlinearity in the predictive regression. We focus on a "feed-forward" neural network architecture, which is formed by an *input* layer, one or more *hidden* layers, and an *output* layer. Basically, in a feed-forward neural network architecture, the *input* layer is formed by the set of predictors, which are interacted with

each other or nonlinearly transformed in the next hidden layers (formed by neurons or nodes). Each *hidden* layer takes predictive signals from the neurons or nodes in the previous *hidden* layer to produce a new signal. Then, these hidden layers are assembled to predict the outcome variable in the *output* layer, which is a linear function that translates the signals from the last *hidden* layer into the forecast of the response variable.

Although neural networks have the advantage of producing a precise approximation of any smooth predictive association (Cybenko, 1989), the researcher must set several tuning parameters when structuring them. In our analysis, we specify two neural networks. The first one has three hidden layers, and the second one has five. Each network contains 32, 16, 8, 4, and 2 neurons, respectively (i.e., a geometric pyramid rule, following Gu et al., 2020 and Christensen et al., 2021). As an activation function, we use the Leaky Rectified Linear Unit (LReLU) function (Maas et al., 2013), which is a refinement of the conventional ReLU.[7] The LReLU is defined as:

$$LReLU(x) = \begin{cases} 0 & if \ \ x < 0 \\ 0.1x & otherwise \end{cases}$$

the LReLU helps avoid the well-known *Dying ReLU problem*, whereby some neurons are essentially *dead* to all inputs (i.e., they remain inactive). This problem affects neural networks' performance.

To estimate the neural network weight parameters, we minimize an objective function based on the mean squared error (MSE) of the predictions for the training sample. To better guard against overfitting, we consider four regularization approaches: learning rate shrinkage, early stopping, dropout (with a rate of 0.8 for each *hidden* layer), and ensembles. Specifically, to optimize the objective function, we use the stochastic gradient descent (SGD) method implemented via the Adam optimizer with a learning rate of 0.001. This method was developed by Kingma and Ba (2014) and has proven to be a computationally efficient algorithm. To initialize the neural network weight parameters, we employ the Glorot normal initializer (Glorot and Bengio, 2010).

Since the stochastic gradient descent (SGD) approach evaluates the gradient from a small random subset of the data, the neural network weight parameters are conditioned by the starting point (i.e., the seed for the random number generator). To overcome this issue and reduce the neural networks' dependency on the seed, we train the architecture 50 times and use different seeds for each training iteration. We assign 20% of the training data available for each forecast to a validation sample. Then, we sort the results by MSE (from lowest to highest) and based on those 50 predictions. Finally, the training of the neural network is based on a set of 500 epochs with an early stopping of 100 (patience).

3.1.4. PCR and PLSR

We also implement Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR). These methods are widely used in settings where there is a relatively small sample size but many, possibly correlated, predictors.

Specifically, based on the linear model specified in (1), PCR and PLSR firstly decompose the matrix of predictors X into orthogonal scores T and loadings P, where $X = TP$. Then, these methods do not regress R on X itself, but they do so on a small subset of the first columns of the scores T. For this reason, PCR and PLSR are known as methods of dimension reduction. The difference between PCR and PLSR is given by the specific decomposition of scores and loadings considered.

PCR considers the scores that are given by the left singular vectors of X, multiplied by the corresponding singular values, and the loadings are the right singular vectors of X. Intuitively, the decomposition implemented by PCR focuses on the variance of X. In contrast, PLSR considers information on both X and R to define the scores and loadings in order to describe as much of the covariance (or the joint variation) between X and R as possible.

To implement these approaches, we employ the pls package in R developed by Mevik and Wehrens (2015). In our analysis, we use the first PCR and PLSR scores.

3.2. Forecasting framework

As mentioned above, we use the sample period beginning in January 1970 and ending in December 2019. Additionally, we choose a forecast horizon of one, three and six months. To simulate a real out-of-sample forecasting exercise, we split the sample into training and testing subsamples and use the former to adjust the models and the latter to assess their predictive accuracy. We begin our empirical analysis by using a training subsample running from January 1970 until December 1999 and a testing subsample beginning in January 2000 and ending in December 2019. Moreover, to implement the forecast combination, we use an initial validation sample (also known as an initial holdout sample) from January 1990 until December 1999 to determine the combining weights. It is important to highlight that to train the models and produce forecasts of the gold risk premium at month $t+1, t+3$ and $t+6$, we only use information up until month t.

To make predictions, we use an expanding window of information to train the models. This approach uses all prior observations to update the model estimation when making each new prediction in the testing sample. For example, to forecast the gold risk premium for the first month in the testing sample, we begin by fitting the model using the initial training sample (i.e., all the information prior to the first month in the testing sample). Then, to predict the gold risk premium for the second month in the testing sample, we update the model adjustment using all prior observations, i.e., we estimate the model using the initial training sample plus the information in the

[7] Which sets $ReLU(x) = 0$ *if* $x \leq 0$.

testing sample from the first month (i.e., all information prior to the second month in the testing sample). We proceed in this manner until we obtain forecasts of the gold risk premium for all of the months in the testing sample. Crucially, we reiterate this predictive approach using an expanding training sample for all the methods under evaluation in this article.

As mentioned above, we use a time series cross-validation approach to set some of the tree-based and regularization methods' tuning parameters, as in Gu et al. (2020) and Drobetz et al. (2021). Specifically, we retain the temporal ordering of the data and split the sample into three distinct subsamples: a training sample, a validation sample, and a test sample. We begin our empirical analysis by selecting 28 years for training (from January 1970 until December 1998) and one year for validation (from January 1999 until December 1999). The remaining 20 years were used for the testing sample (from January 2000 until December 2019). We used the training sample to estimate the model for multiple hyperparameter specifications and the validation sample to tune the hyperparameters to values that minimized the MSE (within the validation sample. In order to avoid recursively refitting models for each month (because it is computationally intensive), we tuned the hyperparameters and refit the models once per year using a recursive window for the training sample, i.e., a progressively larger data set and a validation sample of one year. In other words, after calibrating the hyperparameters and estimating the models for a given set of training and validation samples, we produced forecasts for the following 12 months in the testing sample, before re-updating the tuning parameters, model estimations, and repeating the entire process.

3.3. Performance metrics

3.3.1. Statistical metrics for assessing forecast error

To assess the out-of-sample prediction accuracy of the methods under analysis, we report the out-of-sample R^2 (R^2_{OOS} or $OOS-R^2$) of Goyal and Welch (2008) given by:

$$R^2_{OOS} = 1 - \frac{\sum_{s=t^*}^{T^*}(R_s - \widehat{R}_s)^2}{\sum_{s=t^*}^{T^*}\left(R_s - \widehat{R}_s^B\right)^2},$$

where t^* is the start month of the testing subsample, T^* is the end month of the testing subsample, R_s is the actual gold risk premium in month s of the testing period, $\widehat{R}_s$ is the predicted gold risk premium for month s of the testing period using the specific model under evaluation, and $\widehat{R}_s^B$ is the predicted gold risk premium for month s of the testing period using the benchmark model (the historical mean up until month $s-1$). The method with the *highest* out-of-sample R^2 is deemed the best according to that measure of forecast error. We use the predictability test of Clark and West (2007) to statistically assess whether the method under evaluation provides better prediction accuracy than the historical mean. Formally, the CW test evaluates the null hypothesis $H_0 : R^2_{OOS} \leq 0$ against the $H_1 : R^2_{OOS} > 0$. A rejection of H_0 provides evidence favoring predictability of the tested model against the benchmark model.

3.3.2. Portfolio metrics

Prior literature has shown that models with low statistical performance may still produce economic gains (see, e.g., Kandel and Stambaugh, 1996; and Cenesizoglu and Timmermann, 2012). We evaluate the economic gains expectable from the gold risk premium predictions of several models by building a portfolio that splits wealth between gold, the S&P500 index and a risk-free asset in a recursive fashion. Consider the case of an investor who allocates their wealth between a risk-free asset and two risky assets (gold and the S&P 500 index) by solving a mean-variance problem:

$$\max_{\omega_{t+1}} E_t[r^p_{t+1}] - \frac{1}{2}\gamma Var_t[r^p_{t+1}]$$

whose solution is a dynamically rebalanced portfolio with time-varying weights given by:

$$\omega^*_{t+1} = \frac{1}{\gamma}\Sigma^{-1}E_t[\mu_{t+1}].$$

where Σ is the historical covariance matrix of the risky assets and $E_t[\mu_{t+1}]$ are predicted returns of risky assets. The expected gold returns are based on the predictions from each of the models under evaluation, while the expected return of the S&P 500 is computed as the historical mean in the last 20 months. After calculating the optimal portfolio weights using the different ML methods to estimate the expected returns of gold, we compute several ex-post portfolio performance metrics: the Sharpe ratio (SR), the certainty equivalent returns (CER), the downside risk (DR), the maximum drawdown (MDD), the portfolio turnover (Turn), and a utility-based metric called the performance fee (Δ) (see, e.g., Çakmaklı and van Dijk, 2016). Assuming $\{r^p_{t+1}\}_{t_1}^{T}$ is the time-series of ex-post portfolio returns, the SR is defined as:

$$\widehat{SR}_p = \widehat{\mu}_p / \widehat{\sigma}_p$$

where $\widehat{\mu}_p$ and $\widehat{\sigma}_p$, computed over the OOS period of length P, are the portfolio's mean and standard deviation, respectively. The certainty equivalent return is:

$$\widehat{CER}_p = \widehat{\mu}_p - \frac{1}{\gamma}\widehat{\sigma}_p$$

We also report the downside risk (DR) of the portfolio, defined as

$$DR_p = \left(\frac{1}{T}\sum_{t=1}^{T} -1 * \min\left(R_t^p, 0\right)^2\right)^{0.5},$$

The maximum drawdown (MDD), defined as

$$MDD_p = -1 * min\left(\frac{f\left(1+R_t^p\right)}{g\left(f\left(1+R_t^p\right)\right)} - 1\right)$$

where $f(\bullet)$ is the cumulative product and $g(\bullet)$ is the maximum. Finally, we also report turnover (Turn.) which captures the rebalancing of the portfolio through time. Turnover is defined as

$$Turn_p = \frac{1}{T-M}\sum_{t=1}^{T-M}\sum_{j=1}^{N}\left(\left|\widehat{w}_{j,t+1} - \widehat{w}_{j,t^+}\right|\right)$$

where $\widehat{w}_{k,j,t}$ is the portfolio weight of asset j at time t, $\widehat{w}_{j,t^+}$ is the portfolio weight before rebalancing at $t+1$; and $\widehat{w}_{j,t+1}$ is the desired portfolio weight at time $t+1$, after rebalancing.

The performance fee is the value of Δ that solves $\overline{U}_p(R^p_{ML\_model} - \Delta) = \overline{U}_p(R^p_{benchmark})$, where the average realized utility for an investor with initial wealth W is:

$$\overline{U}_p(R^p) = \frac{W}{P}\sum_{t=0}^{P-1}\left(R^P_{t+1} - \frac{1}{2}\frac{\gamma}{(1+\gamma)}\frac{1}{P}\sum_{t=0}^{P-1}\left(R^P_{t+1} - \overline{R}^P_{t+1}\right)^2\right),$$

where R^P_{t+1} is equal to $(1 + r^P_{t+1} - c_{t+1})$ and represents the return on the optimal portfolio, net of transaction costs, c_{t+1}. Transaction costs are a fixed proportion, c, of the invested wealth, so the overall cost of rebalancing between t and $t+1$ is given by $c_{t+1} = 2c|\omega_{t+1} - \omega_t|$. A positive Δ value indicates that a given investment strategy based on ML methods outperforms a strategy using the benchmark model. We set the relative risk aversion coefficient, γ, to six, and consider three alternative levels of fixed transaction costs, c: 0, 0.1, and 0.3 percent.

4. Baseline results

This section describes our empirical results. We first report forecasting results obtained using the ML methods. Then we provide the results of the forecasting exercises using predictors individually rather all together.

Table 1
OOS-R^2 Values for Forecasts using Several ML Methods.

Method	Overall		Expansion		Recession	
	OOS-R^2	*p*-value	OOS-R^2	*p*-value	OOS-R^2	*p*-value
Ridge	−8.5	0.84	−6.70	0.63	−15.2	0.94
Lasso	−5.8	0.39	−4.66	0.24	−10.2	0.86
ENet	−4.8	0.32	−3.60	0.19	−9.5	0.87
RF	−6.9	0.65	−4.55	0.40	−16.1	0.80
GBRT	−14.4	0.88	−15.29	0.86	−11.0	0.68
NN5	−0.2	0.47	−0.08	0.40	−0.8	0.67
NN3	0.0	0.39	0.09	0.34	−0.3	0.56
Pooled-Avg Methods	−1.7	0.68	−1.04	0.47	−4.3	0.87
Pooled-Avg	−0.25	0.92	−0.07	0.66	−0.93	0.96
PCR	0.02	0.16	0.02	0.19	0.02	0.31
PLSR	0.56	0.07	0.31	0.04	1.50	0.19

Notes: The table reports forecasting performance values for models' one-month ahead predictions of the gold risk premium. It shows the out-of-sample R^2 metric of Goyal and Welch (2008) and the *p*-value from the predictability test of Clark and West (2007), with the null hypothesis being equal forecasting ability. A rejection of the null favors the forecast from the machine learning (ML) model over that of the benchmark model. The table reports results from regularization methods (Ridge, Lasso, and Elastic Net), a Random Forest method (RF), a Gradient Boosting Regression Tree method (GBRT), and Neural Networks (NN) with five and three hidden layers, respectively. "Pooled-Avg Methods" corresponds to a forecast combination model based on all the ML methods under consideration and "Pooled-Avg" is the forecast combination obtained by combining all the individual predictors. PCR refers to the principal component regression method and PLSR stands for the partial least square regression method. See details in section 3.1.3. "Recession" ("Expansion") refers to recession (non-recession) periods in the U.S. according to the NBER.

4.1. Machine learning algorithm results

Table 1 reports our baseline results. For each ML algorithm, it shows the estimated $OOS - R^2$ measure introduced by Goyal and Welch (2008) to evaluate stock return predictability. On the one hand, a positive value on this metric indicates that the model under evaluation produces better OOS predictions than the benchmark model, i.e., the historical mean (we use the same benchmark model as Baur et al. (2020), Dichtl (2020) and Nguyen et al. (2019)). On the other hand, a negative value indicates that the benchmark model outperforms the competing model. We report the *p*-value of the forecasting accuracy test of Clark and West (2007). A rejection of the null hypothesis of equal forecasting ability indicates that the benchmark is outperformed by the model under evaluation. We also report separate results for periods of expansion and recession, defined according to the NBER's business cycle information for the U.S. This stratification is motivated by prior evidence in the stock return predictability literature (see, e.g, Rapach and Zhou, 2013; Neely et al., 2014) and the gold predictability literature (see, e.g., Nguyen et al., 2019; and Dichtl, 2020) showing that models' forecasting ability varies across the business cycle.

Our results show that the $OOS - R^2$ metric is negative across all the regularization and tree methods we evaluated, indicating that none of them outperform the benchmark model out-of-sample. For example, the estimated $OOS - R^2$ of the LASSO method is -5.8 percent, while it is -6.9 percent for the random forest method. The null hypothesis of equal predictive power (as compared to that of the benchmark model) cannot be rejected in any of these cases. These results are consistent with the negative $OOS - R^2$ estimates. In the case of the NN models, we obtain a negative $OOS - R^2$ for the NN5 model, and an estimate of zero for the NN3 model. As with the previous models, none of the NN models deliver superior forecasting accuracy than the benchmark model. When combining the forecasts from all the methods (Pooled-Avg Methods) or from all the individual predictors (Pooled-Avg), we observe slightly better $OOS - R^2$ point estimates than those from the individual models (-1.7 and -0.25). However, these forecasts are still worse than the benchmark model. We obtain a similar result for the PCR method: the predictability gains relative to the benchmark are statistically zero ($OOS - R^2$ of 0.02% and *p*-value of 0.16). Finally, the PLSR method provides the best forecasting performance, with an estimated $OOS - R^2$ of 0.56%. The Clark and West test of equal forecasting accuracy offers statistical confirmation that the PLSR outperforms the benchmark model (*p*-value of 0.07).

Similar results are observed across the different phases of the business cycle: none of the regularization and tree methods produce positive and statistically significant $OOS - R^2$ values; the NN methods produce less negative but still not statistically significant R^2 values, except in the case of the NN3 model during periods of expansion ($OOS - R^2$ of 0.09). Combination forecasts of both methods and individual predictors are unable to beat the benchmark model either, both during periods of expansion and recession. The PCR method delivers the same $OOS - R^2$ estimates across business cycle phases as in the full sample (0.02%), but again from a statistical point of view, it cannot outperform the benchmark model. Finally, the PLSR method produces positive $OOS - R^2$ values (0.31% during expansions and 1.5% during recessions) but these are only statistically significant during periods of expansion. This evidence is consistent with findings by Zhang et al. (2020), who evaluate the ability of ML methods to forecast stock returns and find that in most cases ML methods underperform when compared with the historical mean.

Table 2 shows the portfolio evaluation results. Recall that our portfolio exercise consists in combining gold with a broader stock index (the S&P 500 index) and a risk-free asset. Given the statistical evaluations reported above, we would expect a weak performance of most of the evaluated ML models. However, prior literature has shown that an asset allocation evaluation may differ from a pure statistical evaluation (see, e.g, Kandel and Stambaugh, 1996). Overall, our portfolio results show mixed results as in several cases the ML methods produce higher portfolio performance than the benchmark model, especially when we compare Sharpe ratios, however, in most of the evaluated cases these differences become negligible when transaction costs are considered.

For regularization methods, we observe that both the LASSO model and the Elastic Net model produce higher CER estimates than the benchmark model (3.48 and 3.58 vs. 2.0). As concerns their SR, the three regularization methods outperform the benchmark method (0.18, 0.45, 0.46 vs. 0.12). In the case of the random forest model, we observe a CER of 3.68%, while the GBRT model produces a CER of 0.84%. Thus, whereas the former significantly outperforms the benchmark model, the latter delivers lower performance. Sharpe ratio estimates for these two tree methods (0.41 and 0.17) are both superior to the benchmark estimate of 0.12.

As compared to the results from the regularization and tree methods, those from the NN are mixed. In terms of their CER estimates, we observe that the two NN structures we evaluated yielded worse portfolios than those based on the benchmark model. Nevertheless, they had higher SR estimates. Forecast combinations also offered weak performance. Only the SR of the combination of methods (Pooled-Avg Methods) was higher than the benchmark model. Additionally, the benchmark model outperformed the combination of individual predictors (Pooled-Avg). Among the principal component methods, the PCR and PLSR methods delivered superior SRs than the benchmark model, but they yielded lower CER estimates.

The portfolio risk measures show that most of the evaluated methods do not reduce risk as compared with the benchmark model. Downside risk and maximum drawdown estimates are lowest for the benchmark (2.5% and 13.5%, respectively). For example, the GBRT method produces estimates of 6.3% and 30.9%, for the same metrics, indicating that portfolio returns backed by this ML method are consistently riskier than the benchmark portfolio. Worse results are obtained for the random forest method. Much better results are obtained using the NN models. Indeed, the two evaluated NN frameworks reduce portfolio risk as compared with benchmark model. The NN3 model produces downside risk and maximum drawdown estimates of 0.68% and 2.8%, which are both lower than the benchmark model's metrics. Similarly, the estimates for the NN5 method are 1.38% and 3.85%. Both principal component methods produce slightly higher portfolio risk measures than the benchmark model, with estimates of 2.5% and 14.0% for the PCR method, and 2.6% and 14.3% for the PLSR method, respectively. The turnover metrics also show that the benchmark model produces significantly lower rebalancing than all the methods under evaluation. The significantly lower turnover of the benchmark method helps explain why

Table 2
Portfolio Evaluation for Forecasts based on Several ML Methods.

Method	CER	SR	DR	MDD	Turnover	Δ (0.00)	Δ (0.01)	Δ (0.03)
Benchmark	2.00	0.12	2.49	13.58	0.64			
Ridge	0.84	0.18	5.75	24.09	34.54	−57	−65	−186
Lasso	3.48	0.45	5.30	20.02	30.08	186	113	−13
ENet	3.58	0.46	5.13	18.43	29.27	202	125	3
RF	3.68	0.41	6.87	35.91	32.33	124	−21	−160
GBRT	0.84	0.17	6.25	30.90	40.92	−91	−149	−321
NN5	0.65	0.39	1.38	3.85	6.79	75	46	15
NN3	0.36	0.52	0.68	2.80	3.79	56	46	23
Pooled-Avg Methods	1.15	0.22	5.52	22.26	29.57	−17	1	−102
Pooled-Avg	−0.51	−0.02	2.61	13.59	3.82	−52	19	4
PCR	0.00	0.12	2.53	13.97	0.70	−1	50	48
PLSR	0.71	0.28	2.64	14.34	3.18	61	92	82

Notes: The table reports portfolio metrics for each of the machine learning methods. The table reports results from a benchmark model (the historical mean), regularization methods (Ridge, Lasso, and Elastic Net), a Random Forest method (RF), a Gradient Boosting Regression Tree method (GBRT), and Neural Networks (NN) with five and three hidden layers, respectively. "Pooled-Avg Methods" corresponds to a forecast combination model based on all the ML methods under consideration and "Pooled-Avg" is the forecast combination obtained by combining all the individual predictors. PCR refers to the principal component regression method and PLSR stands for the partial least square regression method. See details in section 3.1.3. CER (%) are the gains/losses in certainty equivalent returns produced by each of the ML methods relative to the benchmark model. SR is the Sharpe ratio; DR is the portfolio downside risk; MDD is the maximum drawdown; Turnover is the portfolio turnover; and Δ(c) is a utility-based performance fee reflecting how much an investor would be willing to pay to switch from the benchmark model to the ML model when building their portfolio. The value of c corresponds to the proportional transaction cost (in percent) per transaction.

this method is hard to beat from an asset allocation perspective.

Finally, the performance fee estimates show that the LASSO and ENet methods deliver positive estimates of 186 and 202 bps, respectively. In the case in which *no* transaction costs are assumed, to build their portfolio, an investor would be willing to pay 186 (202) bps to use the LASSO (ENet) model instead of the historical mean. On the other hand, the Ridge method delivers a negative performance fee estimate of 57 bps. These results are consistent with those based on the model comparisons using CER and SR estimates. Among the tree-based methods, the random forest method produces a performance fee of 124 bps, another indication that this method produces better portfolios than the benchmark mode. Among the remaining methods, the performance fees of the NN methods

Table 3
OOS-R^2 metric for Forecasts based on Individual Predictors.

	Overall		Expansion		Recession	
	OOS-R^2	p-value	OOS-R^2	p-value	OOS-R^2	p-value
dfr	2.1	0.07	−0.9	0.27	13.7	0.07
empl	0.4	0.16	0.7	0.01	−0.5	0.52
hs	1.5	0.02	1.3	0.04	2.3	0.09
ce16ov	0.7	0.09	0.4	0.17	1.8	0.18
unrate	1.5	0.03	0.9	0.09	4.1	0.08
uemplt5	1.3	0.05	−0.3	0.19	7.4	0.04
usgood	0.4	0.14	0.5	0.06	0.1	0.40
uscons	0.6	0.07	0.7	0.03	0.3	0.39
srvprd	0.3	0.23	0.6	0.03	−1.0	0.62
nonrevsl	0.3	0.26	0.2	0.29	0.7	0.01
ppicrm	−0.6	0.70	−1.1	0.85	1.5	0.04
oil	0.2	0.22	1.0	0.04	−3.2	0.90
gprh	0.3	0.10	−0.1	0.22	1.8	0.06
gprht	0.3	0.10	−0.1	0.19	1.6	0.07
gprha	0.1	0.28	−0.1	0.54	0.7	0.09
stocksmom12	0.3	0.14	0.5	0.07	−0.6	0.78
stocksrv1	0.3	0.19	−0.7	0.48	3.9	0.03
goldma29	−0.6	0.16	0.3	0.08	−4.1	0.68
goldma212	0.3	0.06	1.1	0.03	−2.9	0.56
goldma312	−0.6	0.13	0.1	0.08	−3.1	0.59
goldmom9	−1.2	0.30	0.6	0.07	−8.1	0.95
goldmom12	0.1	0.10	1.4	0.03	−4.8	0.80

Notes: The table reports forecasting performance metrics for one-month ahead predictions of the gold risk premium using individual predictors (see online appendix for a definition of each variable). It reports the out-of-sample R^2 metric of Goyal and Welch (2008) and the p-value from the predictability test of Clark and West (2007), with the null hypothesis being equal forecasting ability. A rejection of the null favors the forecast from the machine learning model over that of the benchmark model. "Recession" ("Expansion") refers to recession (non-recession) periods in the U.S. according to the NBER.

are positive (75 bps and 56 bps for the NN3 and NN5 models, respectively) but lower than those of the models mentioned above. Similarly, the performance fee of the PLSR is 61 bps. The GBRT method, the forecast combination methods, and the PCR method are associated with negative performance fees instead. As expected, accounting for transaction costs erodes the performance fee estimates. A proportional performance fee of 10 bps reduces the performance fee of the LASSO and ENet methods to 113 and 125 bps, respectively, while the performance fee of the random forest methods becomes negative. The NN methods and principal component methods produce annualized performance fees of approximately 50 bps. In the most realistic scenario in which proportional transaction costs are considered to be 30 bps, we observe that only the PLSR method produces an economically meaningful performance fee of 82 bps.

Overall, when estimating the gold risk premium, the evidence reported so far suggests that there are no significant statistical gains from using either regularization or tree ML methods in conjunction with a large set of predictors. Forecast combinations of both the ML models and the individual predictors also deliver limited statistical gains as compared to the benchmark method. More positive results were obtained using the PLSR method, especially during periods of economic expansion. From an asset allocation perspective, we determine that the ML methods may outperform the benchmark model, especially when comparing portfolios using the SR and CER estimates. Interestingly, the NN methods appear to help reduce portfolio risk, as proxied by the downside risk and maximum drawdown estimates. Furthermore, the performance fee estimates confirm that the LASSO and ENet models add value from a portfolio perspective when transaction costs are ignored. Net of transaction costs, these gains are considerably smaller. The PLSR is the only method that offers gains that are substantial enough to survive transaction costs to some extent.

4.2. Results for individual predictors

The poor performance of the ML methods so far warrants the implementation of a second empirical exercise in which we evaluate the forecasting accuracy of individual predictors. Due to space constraints, Table 3 only shows the group of predictors which, according to the forecasting accuracy test of Clark and West (2007), demonstrate predictive power either in the whole sample or during one of the phases of the business cycle (expansion/recession). Interestingly, we find that several individual predictors have positive and significant $OOS - R^2$ values. Indeed, several macro-finance variables, geopolitical risk proxies and technical indicators outperform the benchmark model at delivering accurate forecasts of the gold risk premium. The most significant predictor variable is the default spread, with $OOS - R^2$ estimates of 2.1 percent in the overall sample, and 13.7 percent during recession periods. Dichtl (2020) also finds this variable to be a significant predictor of the gold risk premium. Macroeconomic variables such as housing starts and unemployment rates also appear to be significant predictors, since they each have estimated $OOS - R^2$ values of 1.5 percent in the whole sample. As in the case of the default spread, these two macroeconomic variables yield more accurate predictions during recession periods ($OOS - R^2$ of 2.3 and 4.1 percent, respectively). The price of oil also shows forecasting ability only during expansion periods ($OOS - R^2$ of 1.0%), Among the technical indicators, we find that the realized volatility of the stock return volatility has predictive power during periods of expansion ($OOS - R^2$ of 3.9%). Similarly, we find that the gold 12-month moving average, as well as the 9- and 12-month momentum indicators outperform the benchmark model during expansion stages, since they have $OOS - R^2$ values of 1.1, 0.6, and 1.4 percent, respectively.[8]

Table 4 shows the portfolio metrics for these predictors. We observe that several predictors yield superior CER estimates than the benchmark. For example, the default spread is associated with a portfolio that has a SR of 3%, compared to 2% for the benchmark model. Interestingly, the gold return technical indicators deliver the highest portfolio CER estimates, with values ranging from 3.53% (9-month momentum) to 5.54% (12-month moving average). The SR estimates show that all the predictors significantly outperform the benchmark model, with the gold return technical indicators yielding the highest performance according to this metric too. Turning now to the portfolio risk metrics and portfolio turnover, we find that the benchmark model tends to outperform most of the individual predictors, since it produces lower values on these three metrics. Among the full set of predictors, the benchmark model yields the lowest DR estimate, while predictors such as the unemployment rate, production price index, geopolitical indices (gprh and gprht), and stock index return momentum (stockmom) outperform the benchmark model according to the MDD estimates. Finally, the benchmark model delivers the lowest portfolio turnover.

In terms of performance fee estimates, we find that several macroeconomic variables, such as the default spread, housing starts, unemployment rate (unrate), oil price, and geopolitical risk measures, deliver positive performance fees. However, this portfolio gain tends to disappear after controlling for transaction costs. The technical indicators are more interesting in this regard, as we observe that they still deliver economic value net of transaction costs. Predictors such as the gold return momentum indicators and moving average indicators yield significant performance fees of between 100 and 200 bps.

In unreported results (available upon request from the authors), we found that several technical indicators yielded significant

[8] One concern with these results is the possibility that they were obtained by chance, considering that coincidentally around 10 percent (20 out of 186) of the predictors were observed to be statistically significant. Nevertheless, we deem this unlikely because most of the variables that were selected have been identified as significant predictors in other studies too. For example, the default spread has been identified as an important predictor of gold prices by Dichtl (2020), Hollstein et al. (2021), and Pierdzioch et al. (2014); business cycle fluctuations, as captured by the employment/unemployment variables in our set of predictors, have also been highlighted by Aye et al. (2015); inflation variables (the production pricing index, PPI, being our measure) have been identified by Pierdzioch et al. (2014) and Jabeur et al. (2021), among others; oil prices have been identified by Le and Chang (2012) and Tanin et al. (2022); and uncertainty variables, which are closely related to our set of geopolitical risk variables, have been singled out by Baur and Smales (2020), Chiang (2022), Gozgor et al. (2019), Bouoiyour et al. (2018), Balcilar et al. (2016), and Beckmann et al. (2019).

Table 4
Portfolio Evaluation for Forecasts using Individual Predictors.

	CER (%)	SR	DR	MDD	Turnover	Δ (0.00)	Δ (0.01)	Δ (0.03)
Benchmark	2.00	0.12	2.49	13.58	0.64			
dfr	3.00	0.41	5.64	40.56	40.68	147	138	−12
empl	0.55	0.15	5.88	24.25	7.59	−51	80	50
hs	2.22	0.35	5.29	15.57	38.92	103	121	16
ce16ov	2.49	0.34	6.29	25.84	34.88	71	107	−2
unrate	3.36	0.48	4.47	12.90	34.89	217	121	5
uemplt5	3.77	0.46	6.09	23.49	56.88	193	117	−79
usgood	0.40	0.15	4.55	19.82	5.43	−24	74	50
uscons	0.55	0.17	4.83	19.97	8.07	−21	81	47
srvprd	0.74	0.17	6.27	24.68	9.72	−50	69	39
nonrevsl	0.71	0.26	3.25	22.37	13.07	52	72	22
ppicrm	0.21	0.14	3.13	13.05	15.24	−3	9	−50
oil	3.73	0.56	4.65	23.70	41.66	273	129	0
gprh	1.48	0.35	3.55	10.41	9.58	107	75	45
gprht	1.57	0.36	3.60	10.34	10.64	113	73	41
gprha	0.28	0.16	3.28	14.44	4.29	8	56	39
stocksmom12	1.09	0.30	3.18	10.23	5.75	77	56	45
stocksrv1	3.88	0.46	6.36	37.69	51.15	192	112	−21
goldma29	4.54	0.43	8.13	33.43	10.88	109	181	146
goldma212	5.54	0.50	8.09	29.55	7.53	203	247	222
goldma312	4.68	0.44	8.29	38.12	8.37	109	203	175
goldmom9	3.53	0.35	9.20	36.40	11.72	−27	112	79
goldmom12	4.94	0.44	9.16	27.18	12.55	100	158	118

Notes: The table reports portfolio metrics using individual predictors (see online appendix for a definition of each variable). CER(%) are the gains/losses in certainty equivalent returns produced by each of the ML methods relative to the benchmark model. SR is the Sharpe ratio, DR is the portfolio downside risk, MDD is the maximum drawdown, Turnover is the portfolio turnover, and Δ(c) is a utility-based performance fee reflecting how much an investor would be willing to pay to switch from the benchmark model to the ML model when building their portfolio. The value of *c* corresponds to the proportional transaction cost (in percent) per transaction.

Table 5
Robustness (economic restrictions).

Method	Overall		Expansion		Recession			
	OOS-R^2	*p*-value	OOS-R^2	*p*-value	OOS-R^2	*p*-value		
Ridge	−1.95	0.33	−0.86	0.20	−6.16	0.78		
Lasso	−1.43	0.16	0.39	0.07	−8.42	0.78		
ENet	−0.96	0.13	0.83	0.06	−7.85	0.78		
RF	−2.91	0.45	−3.62	0.49	−0.18	0.39		
GBRT	−3.49	0.38	−4.49	0.42	0.36	0.33		
NN5	0.18	0.19	0.33	0.13	−0.39	0.69		
NN3	0.06	0.36	0.08	0.35	−0.05	0.51		
Pooled-Avg Methods	−0.04	0.28	0.36	0.20	−1.54	0.73		
Pooled Avg	−0.18	0.90	−0.08	0.69	−0.55	0.97		
PCR	0.02	0.20	0.03	0.17	−0.01	0.57		
PLSR	0.53	0.09	0.19	0.08	1.80	0.17		
	CER	**SR**	**DR**	**MDD**	**Turnover**	**Δ (0.00)**	**Δ (0.01)**	**Δ (0.03)**
HA	2.00	0.12	2.49	13.58	0.64			
Ridge	1.31	0.23	5.85	24.36	33.82	−16	−1	−114
Lasso	3.00	0.39	5.53	21.96	29.75	131	129	11
ENet	3.10	0.41	5.34	20.41	29.46	148	140	23
RF	3.65	0.40	7.02	34.32	30.13	115	−29	−157
GBRT	1.09	0.19	6.49	31.12	36.88	−78	−96	−247
NN5	0.75	0.29	2.74	14.57	9.00	64	64	30
NN3	−0.04	0.16	1.61	5.35	5.60	11	34	6
Pooled-Avg Methods	1.06	0.20	5.70	20.15	28.11	−31	11	−84
Pooled-Avg	−0.51	−0.02	2.61	13.59	3.82	−52	21	7
PCR	0.00	0.12	2.53	13.97	0.70	−1	50	48
PLSR	0.71	0.28	2.64	14.34	3.18	61	92	82

Notes: The upper panel reports forecasting performance values and the lower panel provides portfolio metric estimates for each of the ML methods considered in the study. As in Campbell and Thompson (2008), the estimated risk premium is restricted to positive values. See the notes for Tables 1 and 2 for model definitions, details on the forecasting test, and portfolio metric definitions.

economic value during periods of expansion when we compared them using SR estimates and performance fees. For example, when the 12-month gold moving average is used as a forecasting signal during periods of expansion, the SR is 0.58, which is around 60% higher than that of the benchmark model. The performance fee estimates show that investors would be willing to pay to use forecasts from the MA and MOM technical indicators instead of the benchmark model to forecast the gold risk premium. During recessions, macroeconomic variables such as the default spread, unemployment rate, aggregate price indicators, and real estate loans outperform the benchmark model according to their SR and CER estimates. However, only the latter two variables deliver significant portfolio gains after accounting for transaction costs.

5. Robustness

In this section, we report results from three additional exercises. First, we investigate whether imposing economic restrictions on the forecasting exercise similarly to the approach taken by Campbell and Thompson (2008) improves the performance of the ML methods. Secondly, we consider longer forecasting horizons (three and six months) than the one evaluated for the baseline results (one month). Thirdly, we combine forecasts to investigate whether this improves prediction accuracy, as in work by Zhang et al. (2020).

5.1. Economically motivated restrictions

Campbell and Thompson (2008) argued that a negative risk premium prediction made no sense for investors as it is inconsistent with theory. These authors proposed to replace negative risk premium predictions with the historical mean and showed that stock return predictability improved. Dichtl (2020) studied how these types of restrictions impact gold forecasting and found that their ability to improve gold predictability is limited.

Table 5 shows the results obtained when replacing negative gold risk premium forecasts with the historical mean. In panel A, we report R^2 estimates for the full sample and periods of expansion and recession, and in panel B, we report portfolio metrics. The results show an improvement in the forecasting ability of the alternative ML methods relative to the benchmark. The LASSO and ENet regularization methods produce positive (0.39% and 0.83%) and statistically significant R^2 values (p-values of 0.07 and 0.06) during periods of expansion. The PLSR method yields the highest performance among the ML models, but unexpectedly, it has slightly worst metrics: the R^2 estimate drops to 0.56% from 0.59% model for the whole sample, while during periods of expansion, it drops from 0.31% to 0.19% (see Table 1). The portfolio metrics in panel B reveal a similar pattern to that observed for the unrestricted case reported in Table 3: most of the ML methods outperform the benchmark model when we compare them using the Sharpe ratio, while only the LASSO, Enet and the random forest method produce higher CER estimates than the benchmark model. In terms of portfolio-related risk measures, the NN3 provides the best performance by producing lower downside risk and maximum drawdown estimates. The benchmark model again produces the lower portfolio turnover among competing models by far. Finally, only in the case of no transaction costs do the LASSO, the ENet, and the random forest methods deliver positive and sizeable performance fees (131 bps, 148 bps, and 115 bps, respectively). When transaction costs are considered, most of the ML methods lose their economic value added. One exception is the PLSR method, which yields a positive performance fee of 82 bps after accounting for transaction costs.

Table 6
Robustness (Long Horizons, OOS forecasting evaluation).

Method	h = 3		h = 6	
	OOS-R^2	p-value	OOS-R^2	p-value
Ridge	−16.03	0.77	−44.32	0.89
Lasso	−16.16	0.76	−55.71	0.90
ENet	−10.09	0.71	−58.08	0.93
RF	−6.72	0.22	−3.29	0.09
GBRT	−19.07	0.66	−23.90	0.30
NN5	−1.87	0.56	−4.78	0.64
NN3	−0.64	0.57	−1.69	0.61
Pooled-Avg Methods	−2.91	0.65	−8.13	0.66
Pooled-Avg	−0.57	0.84	−0.98	0.78
PCR	0.10	0.11	0.21	0.08
PLSR	−0.21	0.41	1.12	0.04

Notes: The table reports forecasting performance values for models' three-month and six-month ahead predictions of the gold risk premium. It shows the out-of-sample R^2 metric of Goyal and Welch (2008) and the p-value from the predictability test of Clark and West (2007), with the null hypothesis being equal forecasting ability. A rejection of the null favors the forecast from the machine learning (ML) model over that of the benchmark model. The table reports regularization methods (Ridge, Lasso and Elastic Net), a Random Forest method (RF), a Gradient Boosting Regression Tree method (GBRT), and Neural Networks method (NN) with 5 and 3 hidden layers, respectively. Pooled-Avg Methods corresponds to a forecast combination model based on all the ML methods under consideration and Pooled-Avg is the forecast combination obtained combining all individual predictors. PCR is principal components regression method and PLSR is partial least square regression method. See details in section 3.1.3. "Recession" ("Expansion") refers to recession (non-recession) periods in the U.S. according to the NBER.

5.2. Long horizons

So far, we have considered a one-month horizon (i.e., our data frequency). We now check the robustness of our results at three- and six-month horizons. Our results are reported in Tables 6 and 7. Table 6 shows that at the three-month horizon none of the ML methods delivers a positive, statistically significant R^2 estimate, indicating that the historical mean produces the best risk premia estimates. The performance of the regularization and tree-based methods is especially poor. Forecast combinations were unable to beat the benchmark model either. A similar pattern is observed at the six-month horizon: most of the ML methods produce negative R^2 estimates. In the case of the random forest methods, the estimated negative R^2 (−3.3) is marginally statistically significant (p-value of 0.09). Interestingly, we observe that the PCR and PLSR methods do produce positive and statistically significant R^2 estimates (0.21% and 1.12%, respectively). This finding confirms our results at the one-month horizon (see Table 1), i.e., that extracting principal components from a large set of predictors seems to yield the most accurate estimations of gold risk premia.

Table 7 presents portfolio metrics at the three-month (upper panel) and six-month (lower panel) horizons. At the three-month horizon, we observe a pattern comparable to that found for the one-month baseline case: (1) the regularization and tree-based methods produced portfolios with higher CER estimates and Sharpe ratios than the benchmark model; (2) the neural networks outperformed the benchmark model when compared using downside risk and maximum drawdown measures; (3) in terms of turnover, the PCR and PLSR methods delivered comparable results to the benchmark model's. The performance fee estimates show that when transaction costs have been accounted for, the potential gains from using the ML models to build portfolios are actually limited. The LASSO method delivers the best performing model, with a performance fee of 120 bps without transaction costs and one of 87 bps when these are included. At the six-month horizon, the portfolio metrics reveal that the estimations of the gold risk premia are poorer than at shorter horizons. Overall, when considering the CER, Sharpe ratios, downside risk, maximum drawdown, and turnover estimates, some of the ML methods outperformed the benchmark model. However, the performance fee estimates demonstrate that significant economic gains are difficult to obtain, even when ignoring transaction costs.

Table 7
Robustness (long horizons, portfolio evaluation).

Method	h = 3							
	CER	SR	DR	MDD	Turnover	Δ (0.00)	Δ (0.01)	Δ (0.03)
HA	0.60	0.07	0.66	3.83	2.27			
Ridge	0.74	0.32	1.49	7.20	21.28	61	56	24
Lasso	1.38	0.48	1.67	9.58	16.26	120	109	87
ENet	1.09	0.40	1.67	9.58	18.83	92	87	62
RF	0.96	0.31	2.43	11.58	20.45	74	63	31
GBRT	0.83	0.30	2.35	11.32	26.89	65	56	20
NN5	0.09	0.31	0.03	0.11	3.89	10	20	13
NN3	−0.06	0.12	0.00	0.00	0.00	−4	16	10
Pooled-Avg Methods	0.55	0.32	0.99	4.00	19.43	49	46	23
Pooled-Avg	−0.04	0.03	0.45	1.53	2.77	−3	28	25
PCR	−0.01	0.05	0.76	4.37	2.22	−2	28	27
PLSR	0.00	0.06	0.78	4.50	2.26	−1	35	33
	h=6							
Method	**CER**	**SR**	**DR**	**MDD**	**Turnover**	**Δ (0.00)**	**Δ (0.01)**	**Δ (0.03)**
HA	0.24	0.08	0.33	1.44	3.45			
Ridge	0.10	0.05	1.37	7.69	15.79	3	−1	−11
Lasso	0.01	0.00	1.39	7.69	26.31	−5	−5	−22
ENet	0.07	0.03	1.39	7.69	26.31	1	3	−14
RF	0.49	0.26	1.28	5.49	19.85	43	39	27
GBRT	0.57	0.29	1.26	5.49	15.79	50	46	36
NN5	0.03	0.05	0.00	0.00	0.00	3	5	3
NN3	0.03	0.11	0.00	0.00	0.00	3	11	10
Pooled-Avg Methods	0.10	0.05	1.28	5.49	16.61	5	15	6
Pooled-Avg	0.01	0.07	0.20	0.88	5.05	1	2	1
PCR	−0.02	0.11	0.41	1.82	3.39	−2	12	11
PLSR	−0.02	0.10	0.44	1.96	3.71	−2	11	9

Notes: The table reports portfolio metrics for each of the ML methods for three-month (h = 3) and six-month (h = 6) investment horizons. The table reports regularization methods (Ridge, Lasso and Elastic Net), a Random Forest method (RF), a Gradient Boosting Regression Tree method (GBRT), and Neural Networks method (NN) with 5 and 3 hidden layers, respectively. Pooled-Avg Methods corresponds to a forecast combination model based on all the ML methods under consideration and Pooled-Avg is the forecast combination obtained combining all individual predictors. PCR is principal components regression method and PLSR is partial least square regression method. See details in section 3.1.3. CER(%) are the gains/losses in certainty equivalent returns produced by each of the ML methods relative to the benchmark model. SR is the Sharpe ratio, DR is the portfolio downside risk, MDD is the maximum drawdown, Turnover is the portfolio turnover, and Δ(c) is a utility-based performance fee reflecting how much an investor would be willing to pay to switch from the benchmark model to the ML model when building their portfolio. The value of c corresponds to the proportional transaction cost (in percent) per transaction.

5.3. Combination forecasts

Previous results considered combination forecasts across methods (Pooled-Avg Methods) and across individual predictors (Pooled-Avg.). Following Zhang et al. (2020), we explore an additional forecast combination scheme incorporating both individual ML methods and the historical average (HA), i.e., the benchmark model. Table 8 shows our results. We find that the combination schemes' $OOS - R^2$ values are superior, but that they still do not outperform the standalone benchmark model in most cases. For example, the $OOS - R^2$ value for the combined LASSO method is −1.2 percent and that for the combined random tree method is −2.1 percent. The *p*-value from the predictability test of Clark and West (2007) also confirms that the benchmark model still yields superior forecasting performance than the combined forecasts. The forecasting accuracy of the NN methods was also slightly better, but the estimated OOS — R2 estimates were small again (−0.04 and 0.04 for the NN5 and NN3 models, respectively). The PLSR is the best performing model among this group of combined forecasts, with a positive and significant $OOS - R^2$ of 0.32. The portfolio evaluation, in panel B of Table 8, also shows that the combined forecasts are limited in their ability to deliver economic value. We observe that the LASSO, ENet, and PLSR methods are the best performing models overall. The portfolio built using the LASSO (ENet) method produces a performance fee of 146 (148) bps without transaction costs and 113 (120) bps with proportional transaction costs of 10 bps. When the transaction cost is set to 30 bps instead, the performance fees are 23 and 30 bps, respectively. The PLSR method produces performance fees of 85 (76) bps with transaction costs of 10 (30) bps. Overall, combining the ML methods with the benchmark model does not produce results that are as promising as those reported by Zhang et al. (2020) in their study using U.S. stocks.

6. Conclusions

This study assesses the accuracy of several machine learning models' predictions of the gold risk premium when using a large set of economic predictors, including both financial and macroeconomic variables on the one hand and a set of technical indicators for stock returns and gold on the other. Our out-of-sample evaluation considers both statistical and economic metrics. We compare the models' forecasting performance with that of the historical mean, our benchmark model.

From a statistical point of view, none of the ML algorithms under evaluation (except for the PLSR method) outperform the benchmark model in our out-of-sample assessments. Similar results are observed when forecast combinations are considered, as well as when examining the models' performance during expansion and recession periods. We do observe slightly better results when evaluating gold risk premium forecasts obtained using individual predictors. Indeed, we find that the default spread, variables relating to the business cycle (housing starts, employment rates, and producer price index), oil prices, and geopolitical risk variables are useful predictors of the gold risk premium during recessions. During periods of expansion, technical indicators such as the 12-month moving average and momentum indicator also have forecasting power according to the statistical metrics.

From an economic point of view, when the gold risk premium estimates are used to build portfolios, even though several of the ML methods (e.g., LASSO, Elastic Net, and Random Forest) outperform the benchmark model according to some performance metrics (e.g.,

Table 8
Robustness (Forecast combination of a ML method and the benchmark model).

Method	Overall		Expansion		Recession			
	OOS-R^2	*p*-value	OOS-R^2	*p*-value	OOS-R^2	*p*-value		
HA + Ridge	−2.95	0.88	−1.92	0.67	−6.91	0.95		
HA + Lasso	−1.18	0.42	−0.25	0.23	−4.77	0.86		
HA + ENet	−0.83	0.35	0.12	0.17	−4.48	0.87		
HA + RF	−2.11	0.66	−0.97	0.43	−6.49	0.80		
HA + GBRT	−4.56	0.89	−4.67	0.87	−4.14	0.70		
HA + NN5	−0.04	0.46	0.05	0.39	−0.37	0.68		
HA + NN3	0.04	0.39	0.08	0.33	−0.13	0.57		
HA + PCR	0.01	0.16	0.01	0.19	0.01	0.31		
HA + PLSR	0.32	0.07	0.20	0.04	0.81	0.20		
	CER	**SR**	**DR**	**MDD**	**Turnover**	**Δ (0.00)**	**Δ (0.01)**	**Δ (0.03)**
HA	2.00	0.12	2.49	13.58	0.64			
HA + Ridge	0.44	0.14	5.27	17.01	27.80	−60	14	−75
HA + Lasso	2.59	0.40	4.69	13.79	26.22	146	113	23
HA + ENet	2.57	0.40	4.58	13.61	25.59	148	120	30
HA + RF	1.82	0.26	6.54	28.02	28.10	−20	−31	−128
HA + GBRT	0.63	0.15	6.13	24.59	34.82	−99	−43	−160
HA + NN5	0.48	0.50	0.75	2.78	4.07	67	69	46
HA + NN3	0.20	0.33	1.03	3.48	2.55	38	43	29
HA + PCR	0.00	0.12	2.51	13.78	0.67	−1	48	46
HA + PLSR	0.54	0.25	2.55	13.96	2.46	48	85	76

Notes: The table's upper panel reports OOS-R^2 values and Clark and West (2007) p-values for assessments of forecasting performance. The lower panel provides portfolio metrics for forecast combinations incorporating the benchmark model (i.e., the historic average, HA) and a specific machine learning model, as in Zhang et al. (2020). See the notes for Tables 1 and 2 for model definitions, details on the forecasting test, and portfolio metric definitions.

the SR and CER estimates), these gains become negligible after accounting for transaction costs. In the case of the individual predictors, we find that several of these deliver higher SR and CER estimates than the benchmark model. Moreover, according to a portfolio metric accounting for transaction costs, several gold return technical indicators deliver significant economic gains net of transaction costs.

Future research could further explore the estimation of the gold risk premium across the business cycle. It would be especially interesting to adopt an ex-ante perspective instead of the ex-post approach taken in this study and in most of the existing literature. This would require the ability to predict economic recessions in real-time (see, e.g., Moench and Stein, 2021; and Gómez-Cram, 2022).

Credit author statement

Juan D. Díaz: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, **Erwin Hansen**: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, **Gabriel Cabrera**: Software, Formal analysis, Investigation.

Notes: The figure plots monthly gold returns (upper panel) and cumulative gold returns (lower panel). The lower panel also reports cumulative S&P 500 returns. The vertical, gray bars represent recession periods according to the National Bureau of Economic Research (NBER). The vertical, dashed line indicates where the sample was split between in-sample and out-of-sample periods.

Declaration of competing interest

There are no financial conflicts of interest to disclose.

Data availability

Data will be made available on request.

Appendix. List of variables and descriptive statistics

Variable Name	Abbrev.	Mean	Std.	Skew.	Kurt.	AC (1)
Dep. Variable						
Gold Excess Return	GOLD	0.24	5.61	0.30	6.74	0.03
Predictors						
Equity Risk Premium	ERP	0.45	4.37	−0.72	5.59	0.04
Dividend Price Ratio (Log)	DP	−3.64	0.42	0.07	1.98	0.99
Dividend Yield (Log)	DY	−3.63	0.42	0.06	2.00	0.99
Earnings Price Ratio (Log)	EP	−2.85	0.46	−0.59	5.16	0.99
Dividend Payout Ratio (Log)	DE	−0.79	0.32	3.13	19.44	0.99
Book-to-Market Ratio	BM	0.48	0.28	0.85	2.50	0.99
Net Equity Expansion	NTIS	0.01	0.02	−0.49	2.76	0.98
US Market Excess Return	MKT	0.56	4.51	−0.56	4.91	0.07
Size Factor	SMB	0.12	3.07	0.57	8.96	0.01
Value Factor	HML	0.31	2.91	0.10	4.87	0.16
Short Term Reversal Factor	STR	0.47	3.19	0.36	8.31	−0.02
Monthly Momentum Factor	MOM	0.62	4.34	−1.30	12.97	0.05
Volatility of the Equity Risk Premium	RVOL	0.15	0.05	0.75	3.43	0.96
T-Bill Rate (Level)	TBL	0.05	0.03	0.61	3.37	0.99
Rel. T-Bill Rate	RTB	0.00	0.01	−0.09	6.27	0.90
Long Term Bond Return	LTY	0.07	0.03	0.49	2.83	0.99
Long Term Rate of Returns	LTR	0.01	0.03	0.38	5.22	0.03
Rel. Bond Rate	RBR	0.00	0.01	−0.11	5.22	0.86
Default Return Spread	DFR	0.00	0.02	−0.42	9.48	−0.06
Term Spread	TMS	0.02	0.01	−0.63	3.58	0.93
Cochrane Piazzesi Factor	CP	0.98	1.51	0.21	3.11	0.82
Effective Federal Funds Rate	FEDFUNDS	−0.01	0.55	−2.32	47.13	0.40
3-Month AA Financial Commercial Paper Rate	CP3MX	−0.01	0.54	−2.19	41.09	0.33
6-Month Treasury Bill	TB6MS	−0.01	0.42	−1.78	24.99	0.35
1-Year Treasury Rate	GS1	−0.01	0.45	−1.44	19.58	0.36
5-Year Treasury Rate	GS5	−0.01	0.34	−0.36	8.98	0.34
3-Month Commercial Paper Minus FEDFUNDS	COMPAPFFX	0.04	0.40	−1.42	13.41	0.76
3-Month Treasury C Minus FEDFUNDS	TB3SMFFM	−0.51	0.71	−2.66	12.81	0.88
6-Month Treasury C Minus FEDFUNDS	TB6SMFFM	−0.39	0.77	−2.70	13.51	0.89
1-Year Treasury C Minus FEDFUNDS	T1YFFM	0.01	0.78	−2.32	12.19	0.86
5-Year Treasury C Minus FEDFUNDS	T5YFFM	0.78	1.42	−1.48	6.43	0.94
10-Year Treasury C Minus FEDFUNDS	T10YFFM	1.16	1.71	−1.29	5.46	0.95
Moody's Aaa Corporate Bond Minus FEDFUNDS	AAAFFM	2.30	1.99	−1.14	4.90	0.96
Moody's Baa Corporate Bond Minus FEDFUNDS	BAAFFM	3.38	2.05	−0.79	4.07	0.96
Default Spread	DFY	1.08	0.44	1.83	7.43	0.96

(*continued on next page*)

(*continued*)

Variable Name	Abbrev.	Mean	Std.	Skew.	Kurt.	AC (1)
Pastor-Stambaugh Liquidity Factor	PS	−0.03	0.06	−1.50	9.28	0.13
Switzerland/U.S. Foreign Exchange Rate	EXSZUSX	−0.25	2.73	−0.09	4.07	0.27
Japan/U.S. Foreign Exchange Rate	EXJPUSX	−0.20	2.59	−0.45	4.23	0.32
U.S./U.K. Foreign Exchange Rate	EXUSUKX	−0.10	2.32	−0.33	4.93	0.34
Canada/U.S. Foreign Exchange Rate	EXCAUSX	0.03	1.44	0.48	9.94	0.27
Inflation Rate, Monthly	INFL	0.00	0.00	−0.05	5.98	0.60
Inflation Rate, YoY	INFA	0.04	0.03	1.38	4.78	0.99
Industrial Production Growth, Monthly	IPM	0.00	0.01	−1.14	8.44	0.35
Industrial Production Growth, YoY	IPA	0.02	0.05	−1.16	5.50	0.97
Housing Starts	HS	0.00	0.08	0.27	3.75	−0.33
M1 Growth, Monthly	M1M	0.00	0.01	1.41	13.44	0.13
M1 Growth, YoY	M1A	0.06	0.04	0.11	3.09	0.97
Orders, Monthly	ORDM	0.00	0.02	−0.20	4.32	−0.05
Orders, YoY	ORDA	0.01	0.07	−1.05	6.06	0.93
Return CRB Spot	CRB	0.00	0.03	−0.55	9.61	0.28
Employment Growth	EMPL	0.00	0.00	−0.52	6.20	0.64
Diffusion Index	DIFF	9.01	18.62	−0.61	3.71	0.86
ISM PMI	PMI	52.52	6.38	−0.55	4.27	0.94
M2 Money Stock	M2SL	0.00	0.00	−0.83	10.74	−0.33
Real M2 Money Stock	M2REAL	0.00	0.00	0.81	6.61	0.60
Monetary Base	BOGMBASE	0.00	0.02	0.59	18.21	−0.06
Total Reserves of Depository Institutions	TOTRESNS	0.00	0.07	1.12	54.11	−0.09
Reserves Of Depository Institutions	NONBORRES	0.00	1.16	3.92	286.94	−0.48
Commercial and Industrial Loans	BUSLOANS	0.00	0.01	−0.56	7.31	−0.29
Real Estate Loans at All Commercial Banks	REALLN	0.00	0.01	−0.58	23.92	−0.34
Total Nonrevolving Credit	NONREVSL	0.00	0.01	0.25	59.58	−0.52
Nonrevolving consumer credit to Personal Income	CONSPI	0.00	0.00	−0.29	26.69	0.07
MZM Money Stock	MZMSL	0.00	0.01	1.29	25.91	−0.04
Consumer Motor Vehicle Loans Outstanding	DTCOLNVHFNM	0.00	0.03	0.07	16.86	−0.43
Total Consumer Loans and Leases Outstanding	DTCTHFNM	0.00	0.02	0.00	85.13	−0.45
Securities in Bank Credit at All Commercial Banks	INVEST	0.00	0.01	−0.58	10.58	−0.33
PPI: Finished Goods	PPIFGS	0.00	0.01	−0.09	7.17	−0.43
PPI: Finished Consumer Goods	PPIFCG	0.00	0.01	−0.12	7.33	−0.43
PPI: Intermediate Materials	PPIITM	0.00	0.01	−0.99	13.33	−0.38
PPI: Crude Materials	PPICRM	0.00	0.04	−0.94	15.37	−0.44
PPI: Metals and metal products:	PPICMM	0.00	0.03	−0.29	6.70	−0.33
Crude Oil, spliced WTI and Cushing	OILPRICEX	0.00	0.10	0.29	19.13	−0.38
CPI: Apparel	CPIAPPSL	0.00	0.01	0.46	6.14	−0.42
CPI: Transportation	CPITRNSL	0.00	0.01	−0.32	10.48	−0.14
CPI: Medical Care	CPIMEDSL	0.00	0.00	−0.17	10.66	−0.50
CPI: Commodities	CUSR0000SAC	0.00	0.01	−0.38	9.22	−0.22
CPI: Durables	CUSR0000SAD	0.00	0.00	0.28	5.76	−0.23
CPI: Services	CUSR0000SAS	0.00	0.00	−1.57	27.83	−0.46
CPI: All Items Less Food	CPIULFSL	0.00	0.00	−0.33	5.88	−0.22
CPI: All items less shelter	CUSR0000SA0L2	0.00	0.00	−0.26	8.16	−0.26
CPI: All items less medical care	CUSR0000SA0L5	0.00	0.00	−0.11	6.39	−0.27
Personal Cons. Expend.: Chain Index	PCEPI	0.00	0.00	−0.19	4.82	−0.27
Personal Cons. Exp: Durable goods	DDURRG3M086SBEA	0.00	0.00	0.06	4.29	−0.39
Personal Cons. Exp: Nondurable goods	DNDGRG3M086SBEA	0.00	0.01	−0.41	7.74	−0.20
Personal Cons. Exp: Services	DSERRG3M086SBEA	0.00	0.00	1.17	17.31	−0.47
Real personal consumption expenditures	DPCERA3M086SBEA	0.00	0.00	−0.16	6.75	−0.20
Real Manu. and Trade Industries Sales	CMRMTSPLX	0.00	0.01	−0.15	4.32	−0.10
Retail and Food Services Sales	RETAILX	0.00	0.01	−0.20	8.60	−0.17
New Orders for Durable Goods	AMDMNOX	0.00	0.04	−0.17	6.42	−0.28
New Orders for Nondefense Capital Goods	ANDENOX	0.00	0.08	−0.01	7.48	−0.39
Unfilled Orders for Durable Goods	AMDMUOX	0.00	0.01	0.64	4.58	0.63
Total Business Inventories	BUSINVX	0.00	0.01	1.75	24.81	0.50
Total Business: Inventories to Sales Ratio	ISRATIOX	0.00	0.02	0.52	7.16	−0.06
Housing Starts, Northeast	HOUSTNE	4.99	0.40	−0.21	3.05	0.86
Housing Starts, Midwest	HOUSTMW	5.51	0.44	−0.75	3.01	0.92
Housing Starts, South	HOUSTS	6.43	0.32	−0.66	3.25	0.95
Housing Starts, West	HOUSTW	5.79	0.39	−1.02	3.77	0.95
New Private Housing Permits (SAAR)	PERMIT	7.19	0.33	−0.78	3.27	0.98
New Private Housing Permits, Northeast (SAAR)	PERMITNE	5.02	0.38	−0.05	2.93	0.92
New Private Housing Permits, Midwest (SAAR)	PERMITMW	5.48	0.40	−0.65	2.71	0.97
New Private Housing Permits, South (SAAR)	PERMITS	6.36	0.33	−0.53	2.91	0.97
New Private Housing Permits, West (SAAR)	PERMITW	5.82	0.39	−0.99	3.77	0.97
Help-Wanted Index for United States	HWI	4.70	182.26	0.10	4.52	−0.32
Ratio of Help Wanted/No. Unemployed	HWIURATIO	0.00	0.03	−0.51	4.97	0.03
Civilian Labor Force	CLF16OV	0.00	0.00	0.33	5.43	−0.16

(*continued on next page*)

(*continued*)

Variable Name	Abbrev.	Mean	Std.	Skew.	Kurt.	AC (1)
Civilian Employment	CE16OV	0.00	0.00	0.00	4.85	0.11
Civilian Unemployment Rate	UNRATE	0.00	0.18	0.54	4.94	0.17
Average Duration of Unemployment (Weeks)	UEMPMEAN	0.02	0.64	−0.10	5.18	−0.10
Civilians Unemployed - Less Than 5 Weeks	UEMPLT5	0.00	0.05	−0.01	4.49	−0.48
Civilians Unemployed for 5–14 Weeks	UEMP5TO14	0.00	0.05	0.33	3.68	−0.26
Civilians Unemployed - 15 Weeks and Over	UEMP15OV	0.00	0.05	0.51	4.53	0.21
Civilians Unemployed for 15–26 Weeks	UEMP15T26	0.00	0.07	−0.05	4.18	−0.11
Civilians Unemployed for 27 Weeks and Over	UEMP27OV	0.00	0.06	0.37	3.86	0.08
Initial Claims	CLAIMSX	0.00	0.05	0.31	5.48	0.02
All Employees: Goods-Producing Industries	USGOOD	0.00	0.00	−1.39	7.83	0.69
All Employees: Mining and Logging: Mining	CES1021000001	0.00	0.02	0.48	60.07	0.16
All Employees: Construction	USCONS	0.00	0.01	−0.30	5.95	0.36
All Employees: Manufacturing	MANEMP	0.00	0.00	−1.69	10.12	0.70
All Employees: Durable goods	DMANEMP	0.00	0.01	−1.70	13.41	0.61
All Employees: Nondurable goods	NDMANEMP	0.00	0.00	−1.16	8.05	0.63
All Employees: Service-Providing Industries	SRVPRD	0.00	0.00	0.17	8.87	0.45
All Employees: Trade, Transportation and Utilities	USTPU	0.00	0.00	−0.43	4.54	0.58
All Employees: Wholesale Trade	USWTRADE	0.00	0.00	−0.49	3.93	0.66
All Employees: Retail Trade	USTRADE	0.00	0.00	0.10	5.11	0.46
All Employees: Financial Activities	USFIRE	0.00	0.00	−0.49	4.24	0.76
All Employees: Government	USGOVT	0.00	0.00	0.88	14.65	0.07
Avg Weekly Hours: Goods-Producing	CES0600000007	40.30	0.67	−0.23	3.05	0.92
Avg Weekly Overtime Hours: Manufacturing	AWOTMAN	0.00	0.13	−0.18	10.44	−0.24
Avg Weekly Hours: Manufacturing	AWHMAN	40.81	0.77	−0.39	3.07	0.95
Avg Hourly Earnings: Goods-Producing	CES0600000008	0.00	0.00	−0.22	11.83	−0.58
Avg Hourly Earnings: Construction	CES2000000008	0.00	0.01	−0.34	10.26	−0.64
Avg Hourly Earnings: Manufacturing	CES3000000008	0.00	0.00	0.25	9.69	−0.55
Real Personal Income	RPI	0.00	0.01	−0.64	24.90	−0.10
Real personal income ex transfer receipts	W875RX1	0.00	0.01	−1.65	33.26	−0.03
IP: Final Products and Nonindustrial Supplies	IPFPNSS	0.00	0.01	−0.57	5.37	0.16
IP: Final Products (Market Group)	IPFINAL	0.00	0.01	−0.40	4.91	0.05
IP: Consumer Goods	IPCONGD	0.00	0.01	−0.10	4.89	−0.05
IP: Durable Consumer Goods	IPDCONGD	0.00	0.02	−0.09	8.28	0.04
IP: Nondurable Consumer Goods	IPNCONGD	0.00	0.01	−0.13	3.24	−0.23
IP: Business Equipment	IPBUSEQ	0.00	0.01	−1.08	8.79	0.22
IP: Materials	IPMAT	0.00	0.01	−1.55	12.05	0.33
IP: Durable Materials	IPDMAT	0.00	0.01	−1.00	7.29	0.49
IP: Nondurable Materials	IPNMAT	0.00	0.01	−1.51	14.99	0.11
IP: Manufacturing (SIC)	IPMANSICS	0.00	0.01	−1.01	7.46	0.37
IP: Residential Utilities	IPB51222S	0.00	0.04	−0.18	4.16	−0.19
IP: Fuels	IPFUELS	0.00	0.02	0.74	10.19	−0.16
Capacity Utilization: Manufacturing	CUMFNS	−0.02	0.62	−0.96	7.07	0.35
VXO	VXOCLSX	19.70	7.31	1.93	9.80	0.85
Caldara-Iacoviello Geopolitical Index A	GPRH	98.70	54.82	2.14	9.99	0.74
Caldara-Iacoviello Geopolitical Index B	GPRHT	101.63	61.18	2.05	9.11	0.75
Caldara-Iacoviello Geopolitical Index C	GPRHA	84.79	58.59	3.74	26.80	0.48
Financial Uncertainty to Circulate A	FINUNC	0.95	0.13	0.65	3.08	0.98
Financial Uncertainty to Circulate B	MACROUNC	0.80	0.10	1.65	5.75	0.99
Financial Uncertainty to Circulate C	REALUNC	0.75	0.06	1.38	4.93	0.98
Moving Average; i = 1, l = 9	STOCKSMA19	0.70	0.46	−0.85	1.72	0.70
Moving Average; i = 1, l = 12	STOCKSMA112	0.73	0.45	−1.01	2.02	0.78
Moving Average; i = 2, l = 9	STOCKSMA29	0.71	0.46	−0.90	1.81	0.75
Moving Average; i = 2, l = 12	STOCKSMA212	0.72	0.45	−0.99	1.98	0.83
Moving Average; i = 3, l = 9	STOCKSMA39	0.71	0.45	−0.93	1.86	0.78
Moving Average; i = 3, l = 12	STOCKSMA312	0.73	0.45	−1.01	2.02	0.83
Momentum; m = 9	STOCKSMOM9	0.72	0.45	−0.98	1.97	0.76
Momentum; m = 12	STOCKSMOM12	0.74	0.44	−1.11	2.22	0.80
Volume; s = 1, l = 9	STOCKSVOL19	0.69	0.46	−0.83	1.69	0.57
Volume; s = 1, l = 12	STOCKSVOL112	0.71	0.45	−0.94	1.88	0.69
Volume; s = 2, l = 9	STOCKSVOL29	0.68	0.47	−0.79	1.63	0.76
Volume; s = 2, l = 12	STOCKSVOL212	0.71	0.46	−0.91	1.83	0.82
Volume; s = 3, l = 9	STOCKSVOL39	0.70	0.46	−0.88	1.77	0.78
Volume; s = 3, l = 12	STOCKSVOL312	0.70	0.46	−0.88	1.78	0.84
Volatility; m = 1	STOCKSRV1	0.48	0.50	0.09	1.01	−0.22
Volatility; m = 3	STOCKSRV3	0.46	0.50	0.16	1.03	0.23
Volatility; m = 6	STOCKSRV6	0.48	0.50	0.08	1.01	0.32
Volatility; m = 9	STOCKSRV9	0.47	0.50	0.11	1.01	0.39
Volatility; m = 12	STOCKSRV12	0.50	0.50	0.01	1.00	0.36
Moving Average; i = 1, l = 9	GOLDMA19	0.58	0.49	−0.34	1.12	0.73
Moving Average; i = 1, l = 12	GOLDMA112	0.59	0.49	−0.36	1.13	0.78

(*continued on next page*)

(*continued*)

Variable Name	Abbrev.	Mean	Std.	Skew.	Kurt.	AC (1)
Moving Average; i = 2, l = 9	GOLDMA29	0.59	0.49	−0.35	1.12	0.80
Moving Average; i = 2, l = 12	GOLDMA212	0.60	0.49	−0.39	1.15	0.85
Moving Average; i = 3, l = 9	GOLDMA39	0.59	0.49	−0.35	1.13	0.82
Moving Average; i = 3, l = 12	GOLDMA312	0.59	0.49	−0.36	1.13	0.86
Momentum; m = 9	GOLDMOM9	0.59	0.49	−0.36	1.13	0.75
Momentum; m = 12	GOLDMOM12	0.59	0.49	−0.37	1.14	0.77
Volatility; m = 1	GOLDRV1	0.48	0.50	0.07	1.00	−0.20
Volatility; m = 3	GOLDRV3	0.47	0.50	0.11	1.01	0.22
Volatility; m = 6	GOLDRV6	0.49	0.50	0.04	1.00	0.35
Volatility; m = 9	GOLDRV9	0.48	0.50	0.07	1.00	0.37
Volatility; m = 12	GOLDRV12	0.48	0.50	0.08	1.01	0.41

Notes: The table reports descriptive statistics for gold excess returns and the set of predictors considered in the forecasting exercise. It reports their mean, standard deviation, skewness, kurtosis, and their first-order autocorrelation (AC(1)).

References

Aye, G., Gupta, R., Hammoudeh, S., Kim, W.J., 2015. Forecasting the price of gold using dynamic model averaging. Int. Rev. Financ. Anal. 41, 257–266.
Amit, Y., Geman, D., 1997. Shape quantization and recognition with randomized trees. Neural Comput. 9 (7), 1545–1588.
Balcilar, M., Gupta, R., Pierdzioch, C., 2016. Does uncertainty move the gold price? New evidence from a nonparametric causality-in-quantiles test. Resour. Pol. 49, 74–80.
Batten, J.A., Lucey, B.M., 2010. Volatility in the gold futures market. Appl. Econ. Lett. 17 (2), 187–190.
Baur, D.G., Dichtl, H., Drobetz, W., Wendt, V.S., 2020. Investing in gold–Market timing or buy-and-hold? Int. Rev. Financ. Anal.
Baur, D.G., Beckmann, J., Czudaj, R., 2016. A melting pot—gold price forecasts under model and parameter uncertainty. Int. Rev. Financ. Anal. 48, 282–291.
Baur, D.G., Smales, L.A., 2020. Hedging geopolitical risk with precious metals. J. Bank. Finance, 105823.
Bekaert, G., Hoerova, M., Xu, N.R., 2020. Risk, Uncertainty and Monetary Policy in a Global World. Available at: SSRN 3599583.
Baker, S.R., Bloom, N., Davis, S.J., 2016. Measuring economic policy uncertainty. Q. J. Econ. 131 (4), 1593–1636.
Beckmann, J., Berger, T., Czudaj, R., 2019. Gold price dynamics and the role of uncertainty. Quant. Finance 19 (4), 663–681.
Bianchi, D., Büchner, M., Tamoni, A., 2020. Bond risk premiums with machine learning. Rev. Financ. Stud.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, CA, USA.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Bouoiyour, J., Selmi, R., Wohar, M.E., 2018. Measuring the response of gold prices to uncertainty: an analysis beyond the mean. Econ. Modell. 75, 105–116.
Cakmaklı, C., van Dijk, D., 2016. Getting the most out of macroeconomic information for predicting excess stock returns. Int. J. Forecast. 32 (3), 650–668.
Caldara, D., Iacoviello, M., 2018. Measuring Geopolitical Risk. FRB International Finance Discussion Paper (1222).
Campbell, J.Y., Thompson, S.B., 2008. Predicting excess stock returns out of sample: can anything beat the historical average? Rev. Financ. Stud. 21 (4), 1509–1531.
Cenesizoglu, T., Timmermann, A., 2012. Do return prediction models add economic value? J. Bank. Finance 36 (11), 2974–2987.
Chiang, T.C., 2022. The effects of economic uncertainty, geopolitical risk and pandemic upheaval on gold prices. Resour. Pol. 76, 102546.
Christiansen, C., Schmeling, M., Schrimpf, A., 2012. A comprehensive look at financial volatility prediction by economic variables. J. Appl. Econom. 27 (6), 956–977.
Christensen, K., Siggaard, M., Veliyev, B., 2021. A Machine Learning Approach to Volatility Forecasting, vol. 3. Department of Economics and Business Economics, Aarhus University.
Clark, T.E., West, K.D., 2007. Approximately normal tests for equal predictive accuracy in nested models. J. Econom. 138 (1), 291–311.
Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems 2 (4), 303–314.
Dichtl, H., 2020. Forecasting excess returns of the gold market: can we learn from stock market predictions? Journal of Commodity Markets 19, 100106.
Drobetz, W., Hollstein, F., Otto, T., Prokopczuk, M., 2021. Estimating Security Betas via Machine Learning. Available at: SSRN 3933048.
Erb, C.B., Harvey, C.R., 2017. The golden constant. J. Invest. 26 (1), 94–100.
Erb, C., Harvey, C.R., Viskanta, T., 2020. Gold, the golden constant, and déjà vu. Financ. Anal. J. 76 (4), 134–142.
Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm. In: Machine Learning: Proceedings of the Thirteenth International Conference, July 3–6, 1996, Bari Italy. Morgan Kaufman, San Francisco, CA, USA, pp. 148–156.
Gambella, C., Ghaddar, B., Naoum-Sawaya, J., 2021. Optimization problems for machine learning: a survey. Eur. J. Oper. Res.
Gargano, A., Timmermann, A., 2014. Forecasting commodity price indexes using macroeconomic and financial predictors. Int. J. Forecast. 30 (3), 825–843.
Gkillas, K., Gupta, R., Pierdzioch, C., 2020. Forecasting realized gold volatility: is there a role of geopolitical risks? Finance Res. Lett. 35, 101280.
Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteen*th International Conference on Artificial Intelligence and Statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings.
Gómez-Cram, R., 2022. Late to recessions: stocks and the business cycle. J. Finance 77 (2), 923–966.
Goyal, A., Welch, I., 2008. A comprehensive look at the empirical performance of equity premium prediction. Rev. Financ. Stud. 21 (4), 1455–1508.
Gozgor, G., Lau, C.K.M., Sheng, X., Yarovaya, L., 2019. The role of uncertainty measures on the returns of gold. Econ. Lett. 185, 108680.
Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Rev. Financ. Stud. 33 (5), 2223–2273.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
Ho, T., 1995. Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition 1, 278–282, 1.
Hoerl, A., Kennard, R., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.
Hollstein, F., Prokopczuk, M., Tharann, B., Simen, C.W., 2021. Predictability in commodity markets: evidence from more than a century. Journal of Commodity Markets 24, 100171.
Huck, N., 2019. Large data sets and machine learning: applications to statistical arbitrage. Eur. J. Oper. Res. 278 (1), 330–342.
Husted, L., Rogers, J., Sun, B., 2020. Monetary policy uncertainty. J. Monet. Econ. 115, 20–36.
Jabeur, S.B., Mefteh-Wali, S., Viviani, J.L., 2021. Forecasting gold price with the XGBoost algorithm and SHAP interaction values. Ann. Oper. Res. 1–21.
Jurado, K., Ludvigson, S.C., Ng, S., 2015. Measuring uncertainty. Am. Econ. Rev. 105 (3), 1177–1216.
Kandel, S., Stambaugh, R.F., 1996. On the predictability of stock returns: an asset allocation perspective. J. Finance 51 (2), 385–424.
Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.
Krauss, C., Do, X.A., Huck, N., 2017. Deep neural networks, gradient-boosted trees, random forests: statistical arbitrage on the S&P 500. Eur. J. Oper. Res. 259 (2), 689–702.
Le, T.H., Chang, Y., 2012. Oil price shocks and gold returns. International Economics 131, 71–103.

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models. Proc. icml 30 (1), 3.
Malliaris, A.G., Malliaris, M., 2016. What drives gold returns? A decision tree analysis. Finance Res. Lett. 13, 45–53.
McCracken, M.W., Ng, S., 2016. FRED-MD: a monthly database for macroeconomic research. J. Bus. Econ. Stat. 34 (4), 574–589.
Mevik, B.H., Wehrens, R., 2015. Introduction to the Pls Package. Help Section of The "Pls" Package of R Studio Software, pp. 1–23.
Moench, E., Stein, T., 2021. Equity Premium Predictability over the Business Cycle. Deutsche Bundesbank Discussion. Paper No. 25/2021.
Neely, C.J., Rapach, D.E., Tu, J., Zhou, G., 2014. Forecasting the equity risk premium: the role of technical indicators. Manag. Sci. 60 (7), 1772–1791.
Nguyen, D.B.B., Prokopczuk, M., Simen, C.W., 2019. The risk premium of gold. J. Int. Money Finance 94, 140–159.
O'Connor, F.A., Lucey, B.M., Batten, J.A., Baur, D.G., 2015. The financial economics of gold—a survey. Int. Rev. Financ. Anal. 41, 186–205.
Pierdzioch, C., Risse, M., Rohloff, S., 2014. On the efficiency of the gold market: results of a real-time forecasting approach. Int. Rev. Financ. Anal. 32, 95–108.
Pierdzioch, C., Risse, M., Rohloff, S., 2015. Forecasting gold-price fluctuations: a real-time boosting approach. Appl. Econ. Lett. 22 (1), 46–50.
Pierdzioch, C., Risse, M., Rohloff, S., 2016a. A boosting approach to forecasting gold and silver returns: economic and statistical forecast evaluation. Appl. Econ. Lett. 23 (5), 347–352.
Pierdzioch, C., Risse, M., Rohloff, S., 2016b. A quantile-boosting approach to forecasting gold returns. N. Am. J. Econ. Finance 35, 38–55.
Pierdzioch, C., Risse, M., 2020. Forecasting precious metal returns with multivariate random forests. Empir. Econ. 58 (3), 1167–1184.
Rapach, D., Zhou, G., 2013. Forecasting stock returns. Handb. Econ. Forecast. 2, 328–383 (Elsevier).
Rapach, D.E., Strauss, J.K., Zhou, G., 2010. Out-of-sample equity premium prediction: combination forecasts and links to the real economy. Rev. Financ. Stud. 23 (2), 821–862.
Risse, M., 2019. Combining wavelet decomposition with machine learning to forecast gold returns. Int. J. Forecast. 35 (2), 601–615.
Schapire, R., 2003. In: Denison, D.D., Hansen, M.H., Holmes, C., Mallick, B., Yu, B. (Eds.), The Boosting Approach to Machine Learning – an Overview. MSRI Workshop on Nonlinear Estimation and Classification. Springer, New York, 2002.
Tanin, T.I., Sarker, A., Brooks, R., Do, H.X., 2022. Does oil impact gold during COVID 19 and three other recent crises? Energy Econ. 108, 105938.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B 58 (1), 267–288. Wiley.
Triki, M.B., Maatoug, A.B., 2020. The Gold Market as a Safe Haven against the Stock Market Uncertainty: Evidence from Geopolitical Risk. Resources Policy, p. 101872.
Weigand, A., 2019. Machine learning in empirical asset pricing. Financ. Mark. Portfolio Manag. 33 (1), 93–104.
Wu, W., Chen, J., Yang, Z., Tindall, M.L., 2020. A cross-sectional machine learning approach for hedge fund return prediction and selection. Manag. Sci.
Zhang, H., He, Q., Jacobsen, B., Jiang, F., 2020. Forecasting stock returns with model uncertainty and parameter instability. J. Appl. Econom.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. B 67 (2), 301–320.