Empirical investigation on supervised machine learning models predicting equity risk premium

Myong Jong Shin[*]

November 16, 2022

Abstract

Using US monthly data, we examine the performance of various supervised machine learning models in forecasting multi-horizon firm-level equity risk premium. Following Gu, Kelly, and Xiu (2020), we use a panel of firm-level US financial characteristics and macroeconomic predictors from March 1957 to December 2016. First, using tests by Giacomini and White (2006), Hansen, Lunde, and Nason (2011), and Hansen (2005), we compare across models and find the model with superior predictive ability. Additionally, we test for conditional superior predictive ability across models using the test by Li, Liao, and Quaedvlieg (2022). Superior predictive ability of a model is conditional on scalar state variable. Among the models tested, we find that generally, partial least squares and random forest show best predictive ability.

Keywords: Big Data, Supervised Machine Learning, Return Predictability, Forecast Evaluation

JEL Codes: C52, C55, C58, G17

[*] Department of Economics, Indiana University Bloomington, 100 S Woodlawn Ave, Bloomington, IN 47405. Email: myonshin@iu.edu

1 Introduction

In empirical finance literature predicting equity risk premium, as stock returns in excess of the risk-free rate, is a very important and popular topic. The field has seen a plethora of models and covariates reported to have good predictability. In this article, we conduct a comparative analysis of supervised machine learning models forecasting multi-horizon individual firm equity risk premium in the US. The goal is to identify models that have good predictability over other models with tests used to check their statistical significance.

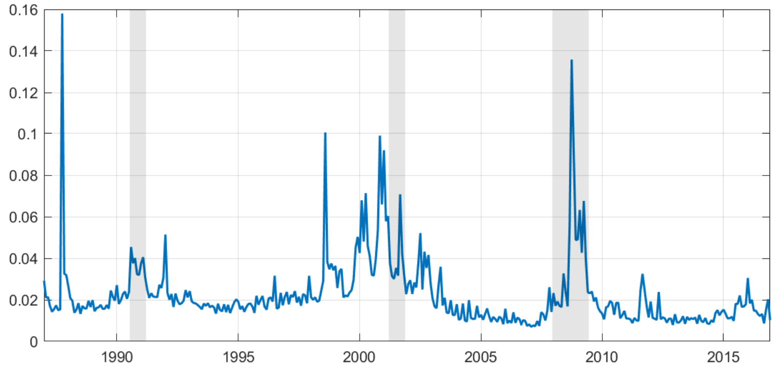

Figure 1: Squared equity risk premium of US stock market.

Given a panel of excess stock returns, for firm i at time t, the naive forecast of zeros for forecasting excess stock returns one month ahead is $\hat{r}_{it} = 0$. Figure 1 shows the squared forecast error from predicting monthly excess stock returns with the naive forecast from March 1987 to December 2016. This is the main benchmark model of predicting individual excess return one month ahead for our study unless mentioned otherwise. It is clear that forecast error is large over important historical events for US financial market such as the stock market crash of 1987, dot com bubble crash of late 90's, great recession of 2008, and many more. Outperforming this naive forecast is not a trivial task, and one of the main goal of this article is to identify forecast models that show smaller forecast error than $\hat{r}_{it} = 0$ from March 1987 to December 2016.

Formally, the benchmark model for forecasting multi-horizon excess returns is Equation 1. h indicates the multi-horizon dimension. In our study we forecast for one month ahead, one quarter ahead, six month ahead, and one year ahead excess returns[1].

$$\sum_{\tau=0}^{h-1} r_{i,t+\tau} = 0 \tag{1}$$

Let $\mathbb{J}$ be the set of models used in this article. For other models, we use various supervised learning methods indexed by $j \in \mathbb{J}$. The naive forecast of zeros is included in $\mathbb{J}$ with $j = 0$ unless mentioned otherwise. Many empirical studies in economics and finance investigate regressions of the form in Equation 2.

$$\sum_{\tau=0}^{h-1} r_{i,t+\tau} = f_j(z_{i,t-1};\theta) + \sum_{\tau=0}^{h-1} \epsilon_{i,t+\tau} \tag{2}$$

For a panel of stocks, firm stocks are indexed by $i = 1, \cdots N_t$ where the number of firm changes per month indexed by $t = 1, \cdots T$. A firm i's h-horizon excess returns is predicted as $\hat{f}_j(z_{i,t-1})$ for model j as the minimizer of an objective function such as least squares. All models use the same covariate $z_{i,t-1}$.

We use squared loss to measure forecast accuracy for all our models[2]. The cross sectional average of squared forecast error from model j is Equation 3.

$$\mathbb{L}_{j,t+h-1} = \frac{1}{N_t} \sum_{i=1}^{N_t} \left[\left(\hat{f}_j(z_{i,t-1}) - \sum_{\tau=0}^{h-1} r_{i,t+\tau} \right)^2 \right] \tag{3}$$

Using Equation 3 for model j, we say that the model j has good predictability over the benchmark unconditionally when it can be shown with statistical significance that the mean squared error of model j is smaller than the benchmark.

$$\mathbb{E}\left[\mathbb{L}_{j,t+h-1} - \mathbb{L}_{0,t+h-1}\right] < 0 \tag{4}$$

We denote $\mathbb{L}_{j,t+h-1} - \mathbb{L}_{0,t+h-1}$ as $\Delta\mathbb{L}_{j,0,t+h-1}$ for simplicity. Figure 2 plots the difference between an alternative model's squared forecast error and that of the benchmark for one month ahead

[1] h=1 for one month ahead, h=3 for one quarter ahead, h=6 for six month ahead, and one year ahead excess returns

[2] There are other choices of loss function that we do not use in this article. A popular alternative to squared loss is Huber loss function. It is a hybrid of squared loss for small errors and absolute loss for relatively large errors.

forecasts $\mathbb{L}_{j,t} - \mathbb{L}_{0,t}, j \in \mathbb{J}$. The figure shows a subsample of January 1997 to December 2006 for readability. During time periods where $\Delta\mathbb{L}_{j,0,t} < 0$ on Figure 2, the model has smaller squared error thus outperforming the benchmark. The alternative models in Figure 2 are Elastic Net model and Random Forest model. Their specification will be elaborated later in Section 3 along with the rest of the supervised machine learning models used in our study[3].

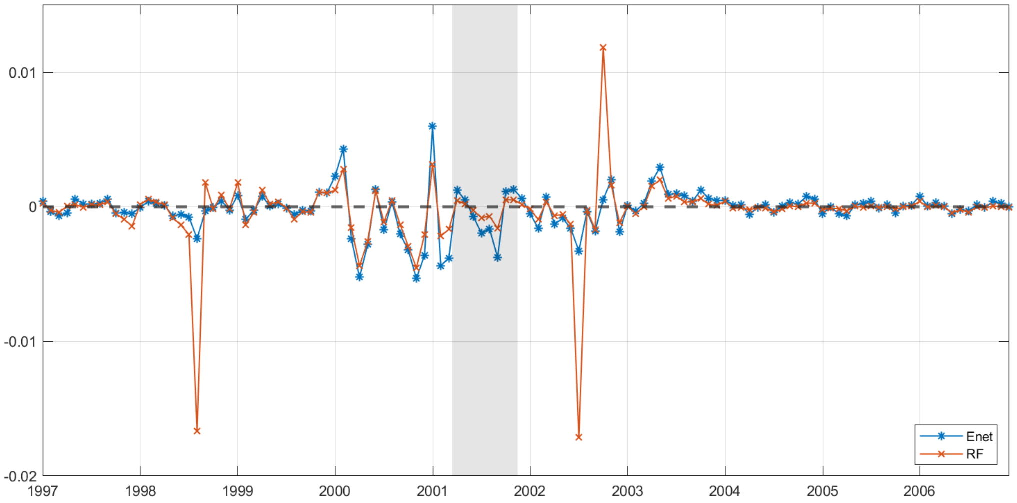

Figure 2: $\Delta\mathbb{L}_{j,0,t}, j \in \mathbb{J}$, January 1997 to December 2006.

The main research questions are twofold. First, for a pair of models $(p, q) \in \mathbb{J}$, we approximate $\mathbb{E}(\Delta\mathbb{L}_{p,q,t+h-1}), p \neq q$ and test significance for all model pairs. Second, for a model $p \in \mathbb{J}$, we approximate the conditional expectation function $\mathbb{E}(\Delta\mathbb{L}_{p,q,t+h-1}), p \neq q$ and test the significance jointly for all $q \in \mathbb{J}$. We call this *all-for-one comparison* of models and we search for models with good predictability over all other models used in this article.

The rest of the paper is organized as follows. Section 2 describes the data and the supervised machine learning models used for the paper for forecasting. Section 3 describes the out-of-sample predictive ability tests used in this paper. Section 4 shows the results and Section 5 concludes. The appendix contains additional results not reported in the article.

[3] The picture of loss difference for the entire testing period of March of 1987 to December of 2016 is provided in the appendix.

2 Data and Models

The key paper that serves as a main motivation for this study is Gu, Kelly, and Xiu (2020)(GKX). Using US financial dataset from Green, Hand, and Zhang (2017) and US macro dataset from Welch and Goyal (2007), GKX predict US firm excess returns one month ahead and check each model's fit out-of-sample. They utilize a wider collection supervised machine learning models compared to the ones used in our study. The contribution of this article is to further their work by extending the forecast exercise to multi-horizon excess returns and perform tests to search for models with good predictability with statistical significance.

2.1 Data

Monthly firm equity returns are from CRSP for all firms listed in NYSE, AMEX, and NASDAQ. Full sample size is from March 1957 to December 2016. Individual excess returns are calculated as individual continuously compounded stock returns minus the risk free rate, and we use 3 month treasury-bill rate as risk free rate. Throughout the period, the number of firms listed per month changes and Figure 3 displays the changes in number of stocks in our sample.

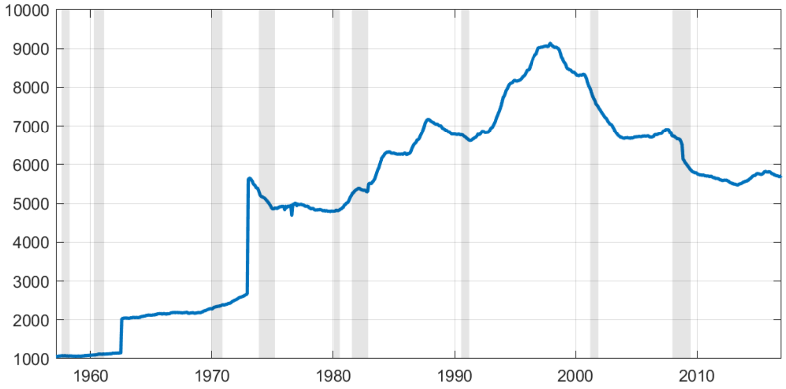

Figure 3: Number of firms N_t.

For predictors, we have three groups of covariates. They are based on those used in GKX and Green, Hand, and Zhang (2017). First, we have 94 panel of firm level characteristics based on the cross-section of stock returns literature. For the panel dataset used in this article, we have over 4 million pooled observations for each firm characteristic[4]. For firm i at time t, $c_{i,t}$ is a 94 by 1

[4] Each firm characteristic is winsorized cross sectionally for 1% and 99%. Refer to the Appendix of Gu, Kelly, and

vector of firm characteristics with $i = 1, \cdots N_t$, $t = 1, \cdots T$.

Second, we have 8 macroeconomic predictors from Welch and Goyal (2007) that are suggested to be good predictors of the market equity premium. They consist of market stock characteristics and bond interest-related predictors[5]. At time t, x_t is a 8 by 1 vector of macroeconomic predictors.

Lastly, we use industry dummies corresponding to the first two digits of US Standard Industrial Classification (SIC) codes. They are used to group firms in similar industries and we have in total 74.

Using three groups, we can construct the vector of features $z_{i,t}$. Let P be the total number of covariates and $z_{i,t}$ be a P by 1 vector. In $z_{i,t}$ we include (1) $c_{i,t}$, (2) the interaction terms between $c_{i,t}$ and x_t, and (3) 74 industry dummies. Therefore $P = 94 * (8 + 1) + 74 = 920$. We use this $z_{i,t}$ to make predictions with all the models in our study using Equation 2.

$$\sum_{\tau=0}^{h-1} r_{i,t+\tau} = f_j(z_{i,t-1};\theta) + \sum_{\tau=0}^{h-1} \epsilon_{i,t+\tau} \tag{2}$$

2.2 Out-of-sample scheme

We split the full sample of March 1957 through December 2016 into three mutually exclusive samples by time. First, training sample, denoted as $\mathbb{T}_1$, is used to estimate the model. Second, validation sample, denoted as $\mathbb{T}_2$, is used to choose hyperparameters of the model if needed. With $\mathbb{T}_1$ and $\mathbb{T}_2$, we can make predictions. Lastly, testing sample, denoted as $\mathbb{T}_3$, is used to compare the predictions with the actual excess returns and compute squared errors.

At the beginning, we have the initial 18 years of training period from March 1957 to December 1974, and 12 years of validation from January 1975 to December 1986. We then forecast for individual excess stock returns for the entire year of 1987, from January to December. In the next step, we roll over our training and validation period for one year and set a new 18 years of training period from March 1958 to December 1975, and 12 years of validation from January 1976 to De-

Xiu (2020) for description of the firm characteristics.

[5] Dividend-price ratio, earning-price ratio, book-to-market ratio, net equity expansion, stock variance, treasury-bill rate, term sprea, and default spread

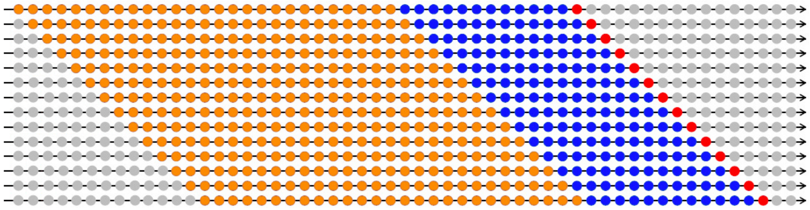

Figure 4: Rolling scheme

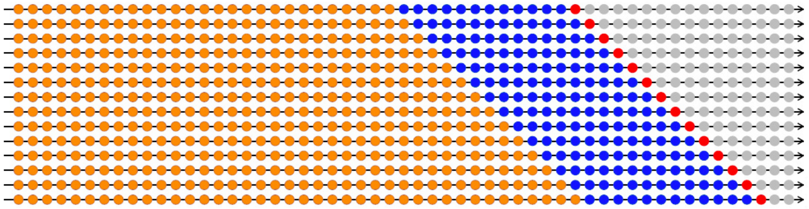

Figure 5: Expanding scheme

cember 1987. We then forecast for individual excess stock returns for the entire year of 1988. We repeat this process until the end of the full sample. In the end we have 30 years of testing period from January 1987 to December 2016 that we can compare with actual excess returns during the same period and calculate forecast errors.

Our way to split data is usually referred to as a rolling scheme. Generally there are two out-of-sample schemes used in econometrics; rolling scheme and expanding scheme. Rolling scheme fixes the sample size of $\mathbb{T}_1$ whereas expanding scheme increases the sample size of $\mathbb{T}_1$ every time we make forecasts. Figure 4 and Figure 5 show the difference between them[6].

2.3 Models

In general terminology used in machine learning literature, prediction problems are divided into unsupervised learning problems and supervised learning problems. Unsupervised learning problems are situations where only the predictors are observable, whereas the latter is a case where we observe both the predictors and the outcome. Thus all of our models in our study are supervised machine learning models. We use them to run regression in Equation 2. In this study we provide only the general description of each model and their tuning parameters if needed. For in depth description of the models there are many good resources. For economists faced with using machine learning

[6] Results using expanding scheme are reported in Appendix for robustness.

models for a prediction problem, Mullainathan and Spiess (2017) and Athey and Imbens (2019) provide excellent recent review of machine learning methods for economists. Books such as James, Witten, Hastie, and Tibshirani (2013) and Efron and Hastie (2021) also discuss machine learning methods in the context of statistics and computer science literature.

2.4 Principal component regression

Principal component regression(PCR) uses principal component(PC) ω_j. PCs are extracted from the sample covariance matrix of Z, and the solution to Equation 5 is the eigenvector associated with jth largest eigenvalue.

$$\omega_j = \arg\max_{\omega} Var(Z\omega), \;\; \omega'\omega = 1, \;\; Cov(Z\omega, Z\omega_k) = 0, \;\; k = 1, 2, \cdots, j-1. \tag{5}$$

For each estimation using training data, we can use scree plot of eigenvalues to pick the number of PCs. This can reduce the dimension and improve forecast. Alternatively, we can use the validation data and choose the number of PCs with smallest mean squared error during the validation period. We find that both approaches provide similar results. For our data a large number of PCs are needed to explain more than 80% of the covariance matrix. However, using too many PCs erodes the benefit of dimension reduction and makes predictions worse. For our dataset we find PCs that explain around 40% of the variance of covariates provide best predictability. The chosen number of PCs is then used for the testing period.

2.5 Partial least squares regression

Partial least squares regression(PLS) extracts PLS components ω_j by solving Equation 6.

$$\omega_j = \arg\max_{\omega} Cov^2(\sum_{\tau=0}^{h-1} r_{t+\tau}, Z\omega), \;\; \omega'\omega = 1, \;\; Cov(Z\omega, Z\omega_k) = 0, \;\; k = 1, 2, \cdots, j-1. \tag{6}$$

Whereas PCR extracts linear combinations of predictors that best explain the covariance matrix, PLS extracts linear combinations of predictors that best explain the squared covariance between the multi-horizon excess returns and the covariates Z. To solve Equation 6 we use SIMPLS algorithm by De Jong (1993). For each estimation using training data, we use the validation data and choose the number of PLS components with smallest mean squared error during the validation period. The chosen number of components is then used for the testing period.

For PCA and PLS, least squares is used to estimate the coefficient vector θ and we use it for prediction $\hat{f}(z_{i,t-1})$. Equation 7 shows the resulting pooled least squares estimator.

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{N_t} \sum_{t=1}^{T} \big(\sum_{\tau=0}^{h-1} r_{i,t+\tau} - f_j(z_{i,t-1};\theta) \big)^2 \tag{7}$$

2.6 LASSO

LASSO stands for least absolute shrinkage and selection operator. The model utilizes a penalty function in addition to the least squares objective function to find the unique optimizer for coefficient vector θ. The LASSO problem can be stated as Equation 8 in lagrangian form with $\lambda \geq 0$.

$$\hat{\theta}_{lasso} = \arg\min_{\theta} \Big(\sum_{i=1}^{N_t} \sum_{t=1}^{T} \big(\sum_{\tau=0}^{h-1} r_{i,t+\tau} - f_j(z_{i,t-1};\theta) \big)^2 + \lambda \sum_{k=1}^{p} |\theta_k| \Big) \tag{8}$$

By constraining θ by $l1$ norm, LASSO induces *sparsity* where we will have some $\hat{\theta}_k = 0$. The optimal tuning parameter λ that controls for the amount of regularization needs to be searched over a grid of candidate values. For each estimation of the model using training data, we use the validation data to fix λ and the chosen value is used for forecasts during testing period.

2.7 Ridge

Ridge model also utilizes a penalty function that constrains the magnitude of θ and its problem can be stated as Equation 9 in lagrangian form with $\lambda \geq 0$.

$$\hat{\theta}_{ridge} = \arg\min_{\theta} \Big(\sum_{i=1}^{N_t} \sum_{t=1}^{T} \big(\sum_{\tau=0}^{h-1} r_{i,t+\tau} - f_j(z_{i,t-1};\theta) \big)^2 + \frac{1}{2} \lambda \sum_{k=1}^{p} \theta_k^2 \Big) \tag{9}$$

In contrast to *variable selection* in LASSO, the Ridge induces *shrinkage* of $\hat{\theta}_k$ near zero. Thus Ridge constrains the magnitude of θ from being too large. The optimal tuning parameter λ that controls for the amount of regularization is determined similar to LASSO.

2.8 Elastic Net

The elastic net model utilizes the penalty functions of both LASSO and ridge as their convex combination. Thus the model estimates θ through both *variable selection* and *shrinkage*. For our study the convex combination parameter ρ is fixed to be 0.5 and the optimal tuning parameter λ is selected using validation data.

$$\hat{\theta}_{enet} = \arg\min_{\theta} \Big(\sum_{i=1}^{N_t} \sum_{t=1}^{T} \big(\sum_{\tau=0}^{h-1} r_{i,t+\tau} - f_j(z_{i,t-1};\theta) \big)^2 + \lambda(1-\rho) \sum_{k=1}^{p} |\theta_k| + \frac{1}{2}\lambda\rho \sum_{k=1}^{p} \theta_k^2 \Big) \tag{10}$$

2.9 Random forest

Random forest is a bootstrap aggregation, or 'bagging' (Breiman (2001)) of individual regression trees. Using an ensemble of B trees, random forest makes a prediction. We use the algorithm by Hastie, Tibshirani, Friedman, and Friedman (2009) for growing random forest. First, using resampled training sample $\{z^b_{i,t-1}, \sum_{\tau=0}^{h-1} r^b_{i,t+\tau}\}, b = 1, \cdots, B$, each regression tree b grows branches in top-to-bottom fashion and at each step, a new branch separates the data from the previous branch binary bins using one of the predictor variables. Its goal is to group observations that are similar to each other. At branch C, a tree chooses a predictor in $z^b_{i,t}$ that can split the data that minimize $l2$ impurity. The loss is defined in Equation 11

$$H(\theta, C) = \frac{1}{|C|} \sum_{z^b_{i,t-1} \in C} \big(\sum_{\tau=0}^{h-1} r^b_{i,t+\tau} - \theta \big)^2 \tag{11}$$

where $|C|$ denotes the number of observations at branch C. We stop branch splitting when the maximum depth of a tree L is reached. With 2^L number of leaves(terminal nodes) we use 300 regressions trees, and each tree has a maximum depth of $L = 6$. The prediction of a tree using bth bootstrap sample $\hat{f}_{tree,b}$ is

$$\hat{f}_{tree,b}(z^b_{i,t-1}, \hat{\theta}^b, L) = \sum_{k=1}^{2^L} \hat{\theta}^b_k \mathbb{1}\{z^b_{i,t-1} \in C_k(L)\}, \ \hat{\theta}^b_k = \frac{1}{|C_k(L)|} \sum_{z^b_{i,t-1} \in C_k(L)} \sum_{\tau=0}^{h-1} r^b_{i,t+\tau} \tag{12}$$

With $\hat{f}_{tree,b}(z^b_{i,t-1}, \hat{\theta}^b, L), b = 1, \cdots, B$, the final random forest prediction $\hat{f}_{rf}$ bags predictions from individual trees as their average.

$$\hat{f}_{rf} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_{tree,b}(z^b_{i,t-1}, \hat{\theta}^b, L) \tag{13}$$

3 Tests

Here we discuss the statistical tests used to answer the main goals of this study. For a pair of models $(p, q) \in \mathbb{J}$, we checked for existence of unit-root for the time series $\Delta\mathbb{L}_{p,q,t+h-1}, p \neq q$ as all the tests in this article assumes stationary of the loss difference. We use unit-root test by Phillips and Perron (1988) to check all model pairs.

3.1 R^2_{OOS} and Equal Predictive Ability Test

Out-of-sample R^2 shows fit during the testing sample period $\mathbb{T}_3$ for model j.

$$R^2_{\text{OOS}} = 1 - \frac{\sum_{(i,t)\in\mathbb{T}_3} \left(\sum_{\tau=0}^{h-1} r_{i,t+\tau} - \hat{f}_j(z_{i,t-1})\right)^2}{\sum_{(i,t)\in\mathbb{T}_3} \left(\sum_{\tau=0}^{h-1} r_{i,t+\tau}\right)^2} \tag{14}$$

R^2_{OOS} is calculated for all models. A positive R^2_{OOS} means that model f_j provides better fit than naive 0 forecast during $\mathbb{T}_3$, and a negative R^2_{OOS} means that model f_j provides worse fit.

For models $(p, q) \in \mathbb{J}, p \neq q$. Giacomini and White (2006) tests the null of equal predictive ability(EPA) of model p and q. Although this does not inform us about which model performs better, this is a standard test used in economics as well as computer science to compare forecast models, so we report the test results.

$$H_0 : \mathbb{E}(\Delta\mathbb{L}_{p,q,t+h-1}) = 0 \tag{15}$$

Rejection of null hypothesis at 5% means that model p and q have different predictive ability with 5% statistical significance. We test EPA for all pairwise combination of models used in our study.

3.2 Superior Predictive Ability and Model confidence set

For models $(p, q) \in \mathbb{J}, p \neq q$ we test the null of superior predictive ability(SPA) of model p over q. We do the test for all possible pairwise combination of models.

$$H_0 : \mathbb{E}(\Delta\mathbb{L}_{p,q,t+h-1}) \leq 0 \tag{16}$$

We also construct a model confidence set(MCS) from Hansen, Lunde, and Nason (2011) with 5% significance level. With MCS we conduct *all-for-one comparison* of models, and we search for models with good predictability over all other models jointly. To define MCS, let d be the loss differences for all pairwise combination of models.

$$d_{p,q,t+h-1} = \Delta\mathbb{L}_{p,q,t+h-1}, \quad \forall p, q \in \mathbb{J} \tag{17}$$

Assuming $\mu_{p,q,h-1} = \mathbb{E}d_{p,q,t+h-1}$ exist and are finite for all model pairs, the set of superior models $\mathbb{M}^*$ is

$$\mathbb{M}^* = \{p \in \mathbb{J} : \mu_{p,q,h-1} \leq 0, \forall q \in \mathbb{J}\} \tag{18}$$

Using the MCS algorithm in Hansen, Lunde, and Nason (2011) we identify the set of superior models jointly in Equation 17 with 5% significance level.

3.3 Conditional Superior Predictive Ability

Lastly, we conduct *all-for-one comparison* of models for Conditional Superior Predictive Ability(CSPA) using the test from Li, Liao, and Quaedvlieg (2022). We test the null of CSPA of model p over all $q \in \mathbb{J}$, given a priori chosen state variable X_t.

$$H_0 : \mathbb{E}(\Delta\mathbb{L}_{p,q,t+h-1} | X_t = x_t) \leq 0, \;\; \forall q \in \mathbb{J}, X_t \in \mathbb{X} \tag{19}$$

The test requires a more stringent condition for rejecting the null as an uniform conditional dominance criterion. The conditional expectation function is approximated via nonparametric series regression of loss difference on the nonparametric basis expansion of the state variable, and the testing algorithm provides the valid critical value for the test. The state variable is scalar.

The purpose of using the test for CSPA is different from the other tests employed in this article. Aside from CSPA, other tests are based on unconditional average performance of the competing models, measured by $\Delta\mathbb{L}_{p,q,t+h-1}$. It integrates out possible heterogeneity across important periods during the testing sample in the data such as economy expansion and recession periods. Although we mainly use CSPA for checking uniform conditional dominance of a model, the conditional evaluation approach can have much potential to find state-dependent *pockets* of varying performance among models in different subsamples. We defer this to future research.

4 Results

Using Equation 2 and various supervised machine learning methods, we forecast different multi-horizon returns using the models described in Section 2. For multi-horizon h, $h = 1$ is one-month-ahead, $h = 3$ is one-quarter-ahead, $h = 6$ is six-month-ahead, and $h = 12$ is one-year-ahead.

Table 1: one-month-ahead results

Rolling	0forecast	Enet	Lasso	PCA40	PLS	Ridge	RF
0forecast		(♠)(⋆)	(♠)(⋆)		(♠)(⋆)	(♠)	
Enet				(♠)	(♠)(⋆)	(♠)	
Lasso				(♠)	(♠)(⋆)	(♠)	
PCA40					(♠)(⋆)	(♠)	
PLS						(♠)	
Ridge							
R^2_{oos}		0.0069	0.0069	0	0.0099	-0.0239	0
$MCS_{5\%}$					✓		

Reject EPA 5% in (♠) under null of $H_0 : \mathbb{E}(\mathbb{L}_t^{col} - \mathbb{L}_t^{row}) = 0$.
Reject SPA 5% in (⋆) under null of $H_0 : \mathbb{E}(\mathbb{L}_t^{col} - \mathbb{L}_t^{row}) \leq 0$.

Table 1 displays test results for $h = 1$, forecasting excess returns one month ahead. We report R^2_{OOS} for individual models and pairwise test results for EPA and SPA, and MCS with 5% significance. Rejection of EPA is marked with (♠), rejection of SPA is marked with (⋆), and models inside MCS is marked with ✓. We can see that only PLS is included in MCS, and also has the highest R^2_{oos}. It rejects SPA null for all models except Ridge. Therefore PLS is preferred model to forecast excess returns.

Table 2: one-quarter-ahead results

Rolling	0forecast	Enet	Lasso	PCA60	PLS	Ridge	RF
0forecast		$(\spadesuit)(\star)$	$(\spadesuit)(\star)$		$(\spadesuit)(\star)$	$(\spadesuit)$	$(\star)$
Enet				$(\spadesuit)$		$(\spadesuit)$	$(\star)$
Lasso				$(\spadesuit)$		$(\spadesuit)$	$(\star)$
PCA60					$(\spadesuit)(\star)$	$(\spadesuit)$	$(\star)$
PLS						$(\spadesuit)$	$(\star)$
Ridge							$(\spadesuit)(\star)$
R^2_{oos}		0.0155	0.0155	0.0022	0.0178	-0.0262	0.0380
$MCS_{5\%}$							✓

Reject EPA 5% in $(\spadesuit)$ under null of $H_0 : \mathbb{E}(\mathbb{L}_t^{col} - \mathbb{L}_t^{row}) = 0$.
Reject SPA 5% in $(\star)$ under null of $H_0 : \mathbb{E}(\mathbb{L}_t^{col} - \mathbb{L}_t^{row}) \leq 0$.

Table 3: six-month-ahead results

Rolling	0forecast	Enet	Lasso	PCA60	PLS	Ridge	RF
0forecast		$(\spadesuit)(\star)$	$(\spadesuit)(\star)$		$(\spadesuit)(\star)$		$(\spadesuit)(\star)$
Enet				$(\spadesuit)$		$(\spadesuit)$	$(\spadesuit)(\star)$
Lasso				$(\spadesuit)$		$(\spadesuit)$	$(\spadesuit)(\star)$
PCA60					$(\spadesuit)(\star)$	$(\spadesuit)$	$(\spadesuit)(\star)$
PLS						$(\spadesuit)$	$(\star)$
Ridge							$(\spadesuit)(\star)$
R^2_{oos}		0.0266	0.0261	0.0034	0.0256	-0.0323	0.0410
$MCS_{5\%}$							✓

Reject EPA 5% in $(\spadesuit)$ under null of $H_0 : \mathbb{E}(\mathbb{L}_t^{col} - \mathbb{L}_t^{row}) = 0$.
Reject SPA 5% in $(\star)$ under null of $H_0 : \mathbb{E}(\mathbb{L}_t^{col} - \mathbb{L}_t^{row}) \leq 0$.

Table 2, 3, and 4 display test results for $h = 3$, $h = 6$, and $h = 12$ respectively. We observe that as the number of horizon h increases, R^2_{OOS} increases for most models and more models are able to be in MCS. Random forest shows best pairwise test results for EPA, SPA, and MCS for all horizons. The Ridge model show poor predictive ability in all our prediction exercises. Because Ridge is the only model that does not reduce the number of parameters estimated, we suspect a repeated exercise in our article using a smaller set of predictors may show changes in performance. We defer this to future research.

Table 4: one-year-ahead results

Rolling	0forecast	Enet	Lasso	PCA60	PLS	Ridge	RF
0forecast		$(\star)$	$(\star)$		$(\star)$		$(\star)$
Enet				(♠)		(♠)	
Lasso				(♠)		(♠)	
PCA60					(♠)$(\star)$		(♠)$(\star)$
PLS						(♠)	
Ridge							(♠)$(\star)$
R^2_{oos}		0.0305	0.0304	0.0044	0.0323	-0.0294	0.0428
$MCS_{5\%}$		✓	✓		✓		✓

Reject EPA 5% in (♠) under null of $H_0 : \mathbb{E}(\mathbb{L}_t^{col} - \mathbb{L}_t^{row}) = 0$.
Reject SPA 5% in $(\star)$ under null of $H_0 : \mathbb{E}(\mathbb{L}_t^{col} - \mathbb{L}_t^{row}) \leq 0$.

Next, we report *all-for-one* test for CSPA. $(\star)$ indicates rejection of CSPA null hypothesis of $\mathbb{E}(\Delta\mathbb{L}_{p,q,t+h-1}|X_t = x_t) \leq 0, \ \forall q \in \mathbb{J}, X_t \in \mathbb{X}$. Model p in red is the benchmark model to beat with all other models. The state variable X_t for CSPA test results is on the first column of each table[7].

Table 5 through Table 8 displays CSPA test results for different forecast horizon. We can see that for all results reported, random forest and PLS are difficult benchmark to reject the null of uniform dominance for all the state variables considered.

Table 5: one-month-ahead, CSPA results

x_t	0forecast	Enet	Lasso	PCA40	PLS	Ridge	RF
ADS		$(\star)$	$(\star)$	$(\star)$		$(\star)$	
ANFCI	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI_credit	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI_leverage		$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI_risk	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
rec_prob	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
vxo	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
gecon	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
mpu	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
epu_comp	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
epu_news	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
JLN_f_h1	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
JLN_m_h1	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
JLN_r_h1	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	

Table 6: one-quarter-ahead, CSPA results

x_t	0forecast	Enet	Lasso	PCA40	PLS	Ridge	RF
ADS	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
ANFCI	$(\star)$			$(\star)$		$(\star)$	
NFCI	$(\star)$		$(\star)$	$(\star)$		$(\star)$	
NFCI_credit	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI_leverage	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI_risk	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
rec_prob	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
vxo	$(\star)$			$(\star)$		$(\star)$	
gecon	$(\star)$			$(\star)$		$(\star)$	
mpu	$(\star)$			$(\star)$		$(\star)$	
epu_comp	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
epu_news	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
JLN_f_h3	$(\star)$			$(\star)$		$(\star)$	
JLN_m_h3	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
JLN_r_h3	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	

[7] Full description of each state variable used can be found in Appendix.

Table 7: six-month-ahead, CSPA results

x_t	0forecast	Enet	Lasso	PCA40	PLS	Ridge	RF
ADS	$(\star)$			$(\star)$		$(\star)$	
ANFCI	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
NFCI	$(\star)$		$(\star)$	$(\star)$		$(\star)$	
NFCI_credit	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
NFCI_leverage	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
NFCI_risk	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
rec_prob	$(\star)$			$(\star)$	$(\star)$	$(\star)$	
vxo	$(\star)$			$(\star)$		$(\star)$	
gecon	$(\star)$			$(\star)$		$(\star)$	
mpu	$(\star)$			$(\star)$		$(\star)$	
epu_comp	$(\star)$			$(\star)$		$(\star)$	
epu_news	$(\star)$			$(\star)$		$(\star)$	

Table 8: one-year-ahead, CSPA results

x_t	0forecast	Enet	Lasso	PCA40	PLS	Ridge	RF
ADS		$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
ANFCI			$(\star)$	$(\star)$		$(\star)$	
NFCI	$(\star)$	$(\star)$		$(\star)$		$(\star)$	
NFCI_credit	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
NFCI_leverage	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
NFCI_risk			$(\star)$	$(\star)$		$(\star)$	
rec_prob		$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
vxo	$(\star)$			$(\star)$	$(\star)$	$(\star)$	
gecon				$(\star)$		$(\star)$	
mpu	$(\star)$			$(\star)$		$(\star)$	
epu_comp	$(\star)$			$(\star)$		$(\star)$	
epu_news	$(\star)$			$(\star)$		$(\star)$	
JLN_f_h12		$(\star)$		$(\star)$	$(\star)$	$(\star)$	
JLN_m_h12		$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
JLN_r_h12	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	

5 Conclusion

We investigate the predictive ability of supervised learning models using covariates from Gu, Kelly, and Xiu (2020) for forecasting firm level US multi-horizon excess returns introduced in recent finance literature using different tests. First, using tests by Giacomini and White (2006), Hansen (2005), and Hansen, Lunde, and Nason (2011), we find that generally partial least squares and random forest perform best in the pairwise and all-for-one model comparisons. Second, using the test of CSPA by Li, Liao, and Quaedvlieg (2022), the results further show that partial least squares and random forest uniformly dominate other models used in this paper.

For future research, we plan to extend the dataset to include more recent data and investigate return predictability different stock portfolios. Additionally, we aim to include more supervised machine learning models and utilize CSPA further to search for state dependency in predictive models' performance in different subsample periods.

6 Appendix

6.1 Additional test results

We forecast different multi-horizon returns using expanding and rolling schemes for robustness, and tables below report results for test for equal predictive ability, SPA, and MCS. h=1 is one-month-ahead, h=3 is one-quarter-ahead, h=6 is six-month-ahead, and h=12 is one-year-ahead.

$$\sum_{\tau=0}^{h-1} r_{i,t+\tau} = f_j(z_{i,t-1};\theta) + \sum_{\tau=0}^{h-1} \epsilon_{i,t+\tau}, j \in \mathbb{J} \tag{20}$$

($\spadesuit$) indicates rejection of EPA test of 5% in under null of $H_0 : \mathbb{E}(\mathbb{L}^{col}_{t+h-1} - \mathbb{L}^{row}_{t+h-1}) = 0$. ($\star$) indicates rejection of SPA test of 5% under null of $H_0 : \mathbb{E}(\mathbb{L}^{col}_{t+h-1} - \mathbb{L}^{row}_{t+h-1}) \leq 0$.

Table 9: one-month-ahead results

Expanding	0forecast	Enet	Lasso	PCA40	PLS	Ridge	RF
0forecast		(♠)(⋆)	(♠)(⋆)		(♠)(⋆)	(♠)	
Enet				(♠)	(♠)(⋆)	(♠)	
Lasso				(♠)	(♠)(⋆)	(♠)	
PCA40					(♠)(⋆)	(♠)	
PLS						(♠)	
Ridge							(⋆)
R^2_{oos}		0.0067	0.0067	0.0011	0.0095	-0.0224	0.0079
$MCS_{5\%}$					✓		✓
Rolling	0forecast	Enet	Lasso	PCA40	PLS	Ridge	RF
0forecast		(♠)(⋆)	(♠)(⋆)		(♠)(⋆)	(♠)	
Enet				(♠)	(♠)(⋆)	(♠)	
Lasso				(♠)	(♠)(⋆)	(♠)	
PCA40					(♠)(⋆)	(♠)	
PLS						(♠)	
Ridge							
R^2_{oos}		0.0069	0.0069	0	0.0099	-0.0239	0
$MCS_{5\%}$					✓		

Table 10: one-quarter-ahead results

Expanding	0forecast	Enet	Lasso	PCA60	PLS	Ridge	RF
0forecast		(♠)(⋆)	(♠)(⋆)		(♠)(⋆)		(♠)(⋆)
Enet				(♠)	(⋆)	(♠)	(⋆)
Lasso				(♠)	(⋆)	(♠)	(⋆)
PCA60					(♠)(⋆)		(♠)(⋆)
PLS						(♠)	(⋆)
Ridge							(♠)(⋆)
R^2_{oos}		0.0152	0.0152	0.0021	0.0178	-0.0184	0.0367
$MCS_{5\%}$							✓
Rolling	0forecast	Enet	Lasso	PCA60	PLS	Ridge	RF
0forecast		(♠)(⋆)	(♠)(⋆)		(♠)(⋆)	(♠)	(⋆)
Enet				(♠)		(♠)	(⋆)
Lasso				(♠)		(♠)	(⋆)
PCA60					(♠)(⋆)	(♠)	(⋆)
PLS						(♠)	(⋆)
Ridge							(♠)(⋆)
R^2_{oos}		0.0155	0.0155	0.0022	0.0178	-0.0262	0.0380
$MCS_{5\%}$							✓

Table 11: six-month-ahead

Expanding	0forecast	Enet	Lasso	PCA60	PLS	Ridge	RF
0forecast		(♠)(⋆)	(♠)(⋆)		(♠)(⋆)		(♠)(⋆)
Enet				(♠)		(♠)	(⋆)
Lasso				(♠)		(♠)	(⋆)
PCA60					(♠)(⋆)		(♠)(⋆)
PLS						(♠)	
Ridge							(♠)(⋆)
R^2_{oos}		0.0241	0.0237	0.0040	0.0266	-0.0128	0.0324
$MCS_{5\%}$		✓	✓		✓		✓
Rolling	0forecast	Enet	Lasso	PCA60	PLS	Ridge	RF
0forecast		(♠)(⋆)	(♠)(⋆)		(♠)(⋆)		(♠)(⋆)
Enet				(♠)		(♠)	(♠)(⋆)
Lasso				(♠)		(♠)	(♠)(⋆)
PCA60					(♠)(⋆)	(♠)	(♠)(⋆)
PLS						(♠)	(⋆)
Ridge							(♠)(⋆)
R^2_{oos}		0.0266	0.0261	0.0034	0.0256	-0.0323	0.0410
$MCS_{5\%}$							✓

Table 12: one-year-ahead

Expanding	0forecast	Enet	Lasso	PCA60	PLS	Ridge	RF
0forecast		(⋆)	(⋆)		(⋆)		(♠)(⋆)
Enet				(♠)	(⋆)	(♠)	
Lasso				(♠)	(⋆)	(♠)	
PCA60					(♠)(⋆)		(♠)(⋆)
PLS						(♠)	
Ridge							(♠)(⋆)
R^2_{oos}		0.0263	0.0266	0.0040	0.0320	-0.0084	0.0338
$MCS_{5\%}$		✓			✓		✓
Rolling	0forecast	Enet	Lasso	PCA60	PLS	Ridge	RF
0forecast		(⋆)	(⋆)		(⋆)		(⋆)
Enet				(♠)		(♠)	
Lasso				(♠)		(♠)	
PCA60					(♠)(⋆)		(♠)(⋆)
PLS						(♠)	
Ridge							(♠)(⋆)
R^2_{oos}		0.0305	0.0304	0.0044	0.0323	-0.0294	0.0428
$MCS_{5\%}$		✓	✓		✓		✓

Tables below report test results for CSPA. $(\star)$ indicates rejection of CSPA 5% of null of H_0 : $\mathbb{E}(\Delta\mathbb{L}_{p,q,t+h-1}|X_t = x_t) \leq 0,\ \forall q \in \mathbb{J}, X_t \in \mathbb{X}$. The state variable X_t used for CSPA test is listed on the first column of every table.

Table 13: one-month-ahead, expanding

x_t	0forecast	Enet	Lasso	PCA40	PLS	Ridge	RF
ADS	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
ANFCI	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI_credit	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI_leverage	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI_risk	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
rec_prob	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
vxo	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
gecon	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
mpu	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
epu_comp	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
epu_news	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
JLN_f_h1	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
JLN_m_h1	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
JLN_r_h1	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	

Table 14: one-month-ahead, rolling

x_t	0forecast	Enet	Lasso	PCA40	PLS	Ridge	RF
ADS		$(\star)$	$(\star)$	$(\star)$		$(\star)$	
ANFCI	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI_credit	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI_leverage		$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI_risk	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
rec_prob	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
vxo	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
gecon	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
mpu	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
epu_comp	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
epu_news	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
JLN_f_h1	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
JLN_m_h1	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
JLN_r_h1	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	

Table 15: one-quarter-ahead, expanding

x_t	0forecast	Enet	Lasso	PCA40	PLS	Ridge	RF
ADS	$(\star)$			$(\star)$		$(\star)$	
ANFCI	$(\star)$			$(\star)$		$(\star)$	
NFCI	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI_credit	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI_leverage	$(\star)$			$(\star)$		$(\star)$	
NFCI_risk	$(\star)$			$(\star)$		$(\star)$	
rec_prob	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
vxo	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
gecon	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
mpu	$(\star)$			$(\star)$		$(\star)$	
epu_comp	$(\star)$			$(\star)$		$(\star)$	
epu_news	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
JLN_f_h3	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
JLN_m_h3	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
JLN_r_h3	$(\star)$			$(\star)$		$(\star)$	

Table 16: one-quarter-ahead, rolling

x_t	0forecast	Enet	Lasso	PCA40	PLS	Ridge	RF
ADS	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
ANFCI	$(\star)$			$(\star)$		$(\star)$	
NFCI	$(\star)$		$(\star)$	$(\star)$		$(\star)$	
NFCI_credit	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI_leverage	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI_risk	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
rec_prob	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
vxo	$(\star)$			$(\star)$		$(\star)$	
gecon	$(\star)$			$(\star)$		$(\star)$	
mpu	$(\star)$			$(\star)$		$(\star)$	
epu_comp	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
epu_news	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
JLN_f_h3	$(\star)$			$(\star)$		$(\star)$	
JLN_m_h3	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
JLN_r_h3	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	

Table 17: six-month-ahead, expanding

x_t	0forecast	Enet	Lasso	PCA40	PLS	Ridge	RF
ADS	$(\star)$			$(\star)$		$(\star)$	
ANFCI	$(\star)$			$(\star)$		$(\star)$	
NFCI	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI_credit	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI_leverage	$(\star)$	$(\star)$		$(\star)$		$(\star)$	$(\star)$
NFCI_risk	$(\star)$			$(\star)$		$(\star)$	
rec_prob	$(\star)$			$(\star)$		$(\star)$	
vxo	$(\star)$			$(\star)$		$(\star)$	
gecon	$(\star)$			$(\star)$		$(\star)$	
mpu	$(\star)$			$(\star)$		$(\star)$	
epu_comp	$(\star)$			$(\star)$		$(\star)$	
epu_news	$(\star)$			$(\star)$		$(\star)$	

Table 18: six-month-ahead, rolling

x_t	0forecast	Enet	Lasso	PCA40	PLS	Ridge	RF
ADS	$(\star)$			$(\star)$		$(\star)$	
ANFCI	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
NFCI	$(\star)$		$(\star)$	$(\star)$		$(\star)$	
NFCI_credit	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
NFCI_leverage	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
NFCI_risk	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
rec_prob	$(\star)$			$(\star)$	$(\star)$	$(\star)$	
vxo	$(\star)$			$(\star)$		$(\star)$	
gecon	$(\star)$			$(\star)$		$(\star)$	
mpu	$(\star)$			$(\star)$		$(\star)$	
epu_comp	$(\star)$			$(\star)$		$(\star)$	
epu_news	$(\star)$			$(\star)$		$(\star)$	

Table 19: one-year-ahead, expanding

x_t	0forecast	Enet	Lasso	PCA40	PLS	Ridge	RF
ADS	$(\star)$			$(\star)$		$(\star)$	
ANFCI				$(\star)$		$(\star)$	
NFCI	$(\star)$			$(\star)$		$(\star)$	
NFCI_credit	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
NFCI_leverage	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	$(\star)$
NFCI_risk				$(\star)$		$(\star)$	
rec_prob	$(\star)$			$(\star)$		$(\star)$	
vxo	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
gecon	$(\star)$			$(\star)$		$(\star)$	
mpu	$(\star)$			$(\star)$		$(\star)$	
epu_comp	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	
epu_news	$(\star)$			$(\star)$		$(\star)$	
JLN_f_h12	$(\star)$			$(\star)$		$(\star)$	
JLN_m_h12		$(\star)$	$(\star)$	$(\star)$		$(\star)$	
JLN_r_h12	$(\star)$	$(\star)$	$(\star)$	$(\star)$		$(\star)$	

Table 20: one-year-ahead, rolling

x_t	0forecast	Enet	Lasso	PCA40	PLS	Ridge	RF
ADS		$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
ANFCI			$(\star)$	$(\star)$		$(\star)$	
NFCI	$(\star)$	$(\star)$		$(\star)$		$(\star)$	
NFCI_credit	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
NFCI_leverage	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
NFCI_risk			$(\star)$	$(\star)$		$(\star)$	
rec_prob		$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
vxo	$(\star)$			$(\star)$	$(\star)$	$(\star)$	
gecon				$(\star)$		$(\star)$	
mpu	$(\star)$			$(\star)$		$(\star)$	
epu_comp	$(\star)$			$(\star)$		$(\star)$	
epu_news	$(\star)$			$(\star)$		$(\star)$	
JLN_f_h12		$(\star)$		$(\star)$	$(\star)$	$(\star)$	
JLN_m_h12		$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	
JLN_r_h12	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	$(\star)$	

6.2 List of variables used

The firm characteristics used in the empirical study are from the data collected in Green, Hand, and Zhang (2017) and Gu, Kelly, and Xiu (2020). The macro time series used are from Welch and Goyal (2007). Table 21 lists the state variable used in testing CSPA. Uncertainty indices JLN(Jurado, Ludvigson, and Ng (2015)) for one-quarter-ahead are not available.

Table 21: List of state variables

x_t	
ADS	Aruoba-Diebold-Scotti index(Philadelphia Fed)
ANFCI	Adjusted national financial conditions index(Chicago Fed)
NFCI	National financial conditions index(Chicago Fed)
NFCI_credit	National Financial Conditions Credit Subindex(Chicago Fed)
NFCI_leverage	National Financial Conditions Leverage Subindex(Chicago Fed)
NFCI_risk	National Financial Conditions Risk Subindex(Chicago Fed)
rec_prob	Smoothed U.S. Recession Probabilities(St Louis FRED)
vxo	Volatility index(CBOE)
gecon	Global Economic condition indicator(Baumeister, Korobilis, and Lee (2022))
mpu	Monetary Policy Uncertainty index(Husted, Rogers, and Sun (2020))
epu_comp	Economic Policy Uncertainty(USA); three component index(Baker, Bloom, and Davis (2016))
epu_news	Economic Policy Uncertainty(USA); news based index(Baker, Bloom, and Davis (2016))
JLN_f	Financial Uncertainty(Jurado, Ludvigson, and Ng (2015))
JLN_m	Macroeconomic Uncertainty(Jurado, Ludvigson, and Ng (2015))
JLN_r	Real Uncertainty(Jurado, Ludvigson, and Ng (2015))

6.3 Loss difference

We report loss difference between individual alternative models and the benchmark of naive 0 forecast, $\mathbb{L}_{p,t} - \mathbb{L}_{0,t}$ for $p \in \mathbb{J}$. Figures display the resulting time series for each horizon h.

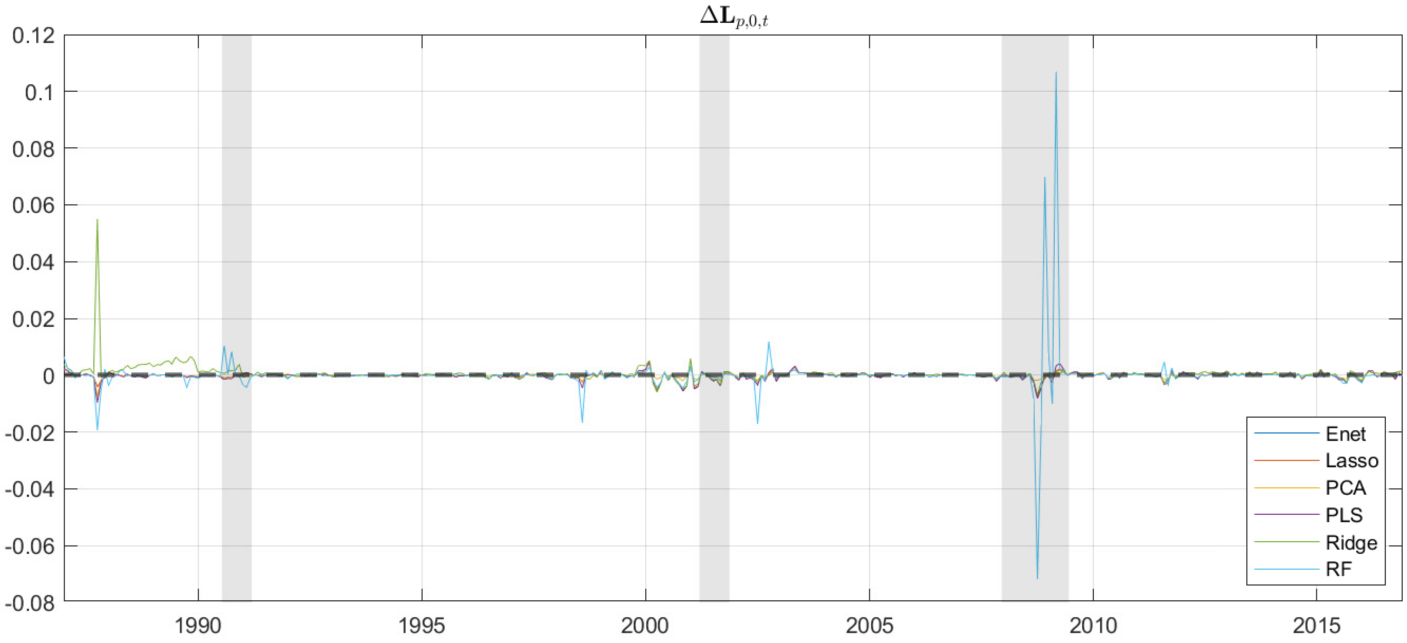

Figure 6: $\mathbb{L}_{p,t} - \mathbb{L}_{0,t}$ for $p \in \mathbb{J}$, one month ahead.

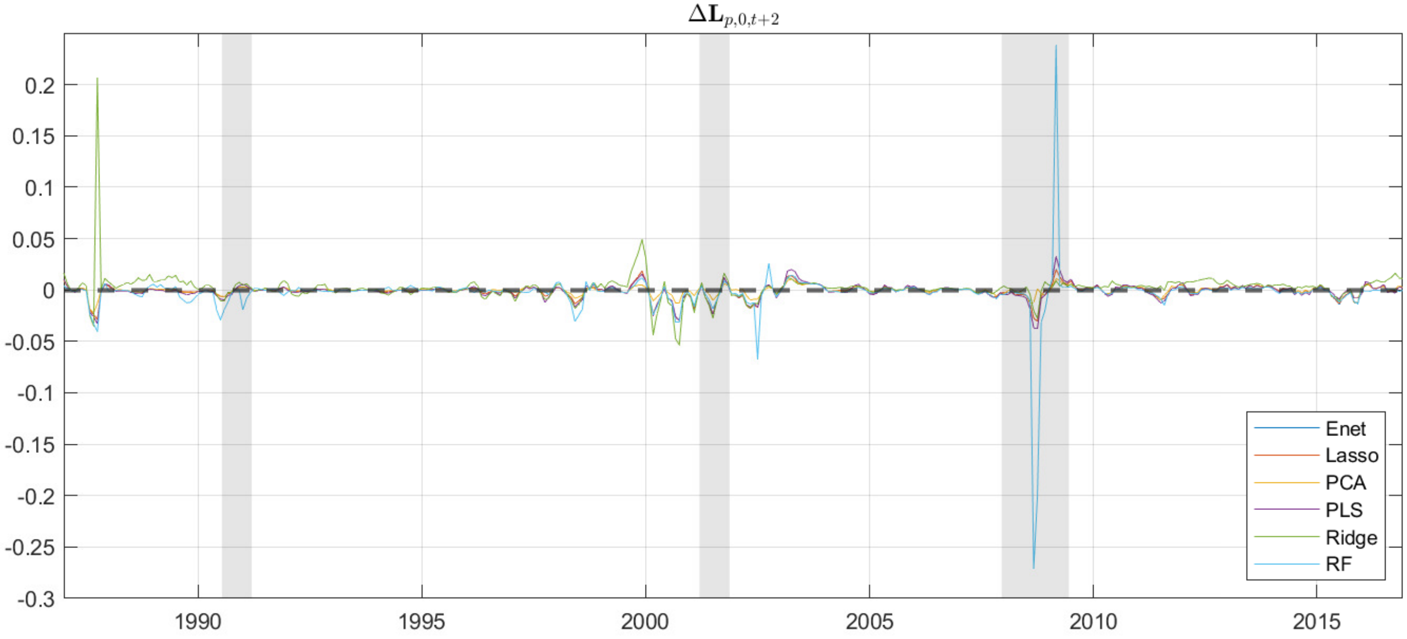

Figure 7: $\mathbb{L}_{p,t+2} - \mathbb{L}_{0,t+2}$ for $p \in \mathbb{J}$, one quarter ahead.

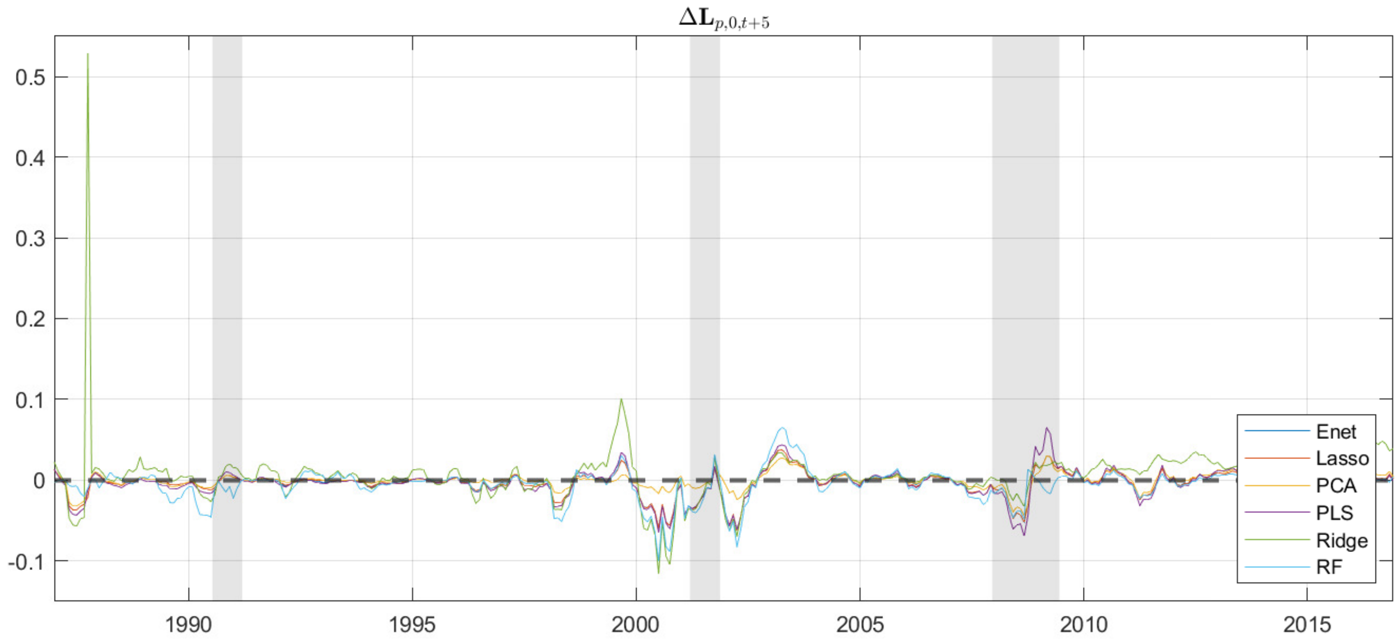

Figure 8: $\mathbb{L}_{p,t+5} - \mathbb{L}_{0,t+5}$ for $p \in \mathbb{J}$, six month ahead.

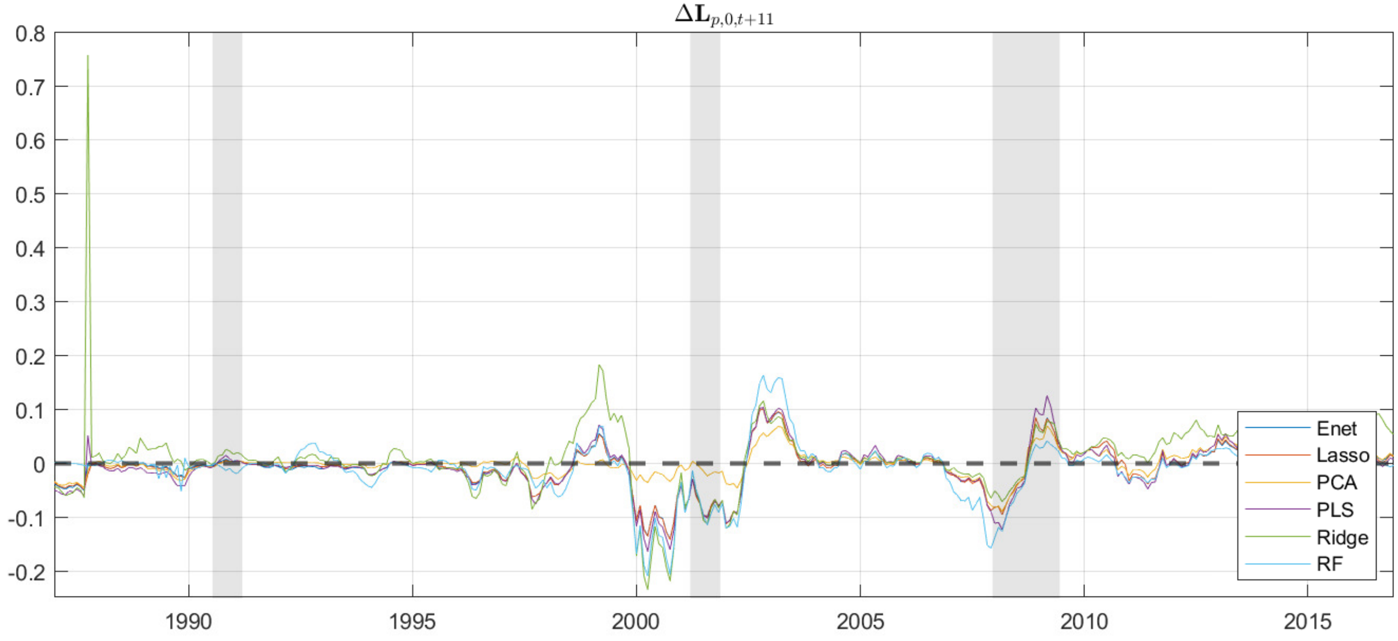

Figure 9: $\mathbb{L}_{p,t+11} - \mathbb{L}_{0,t+11}$ for $p \in \mathbb{J}$, one year ahead.

References

Athey, S., and G. W. Imbens (2019): “Machine learning methods that economists should know about,” *Annual Review of Economics*, 11, 685–725.

Baker, S. R., N. Bloom, and S. J. Davis (2016): “Measuring economic policy uncertainty,” *The quarterly journal of economics*, 131(4), 1593–1636.

Baumeister, C., D. Korobilis, and T. K. Lee (2022): “Energy markets and global economic conditions,” *Review of Economics and Statistics*, 104(4), 828–844.

Breiman, L. (2001): “Random forests,” *Machine learning*, 45(1), 5–32.

De Jong, S. (1993): “SIMPLS: an alternative approach to partial least squares regression,” *Chemometrics and intelligent laboratory systems*, 18(3), 251–263.

Efron, B., and T. Hastie (2021): *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science*, vol. 6. Cambridge University Press.

Giacomini, R., and H. White (2006): “Tests of conditional predictive ability,” *Econometrica*, 74(6), 1545–1578.

Green, J., J. R. Hand, and X. F. Zhang (2017): “The characteristics that provide independent information about average US monthly stock returns,” *The Review of Financial Studies*, 30(12), 4389–4436.

Gu, S., B. Kelly, and D. Xiu (2020): “Empirical asset pricing via machine learning,” *The Review of Financial Studies*, 33(5), 2223–2273.

Hansen, P. R. (2005): “A test for superior predictive ability,” *Journal of Business & Economic Statistics*, 23(4), 365–380.

Hansen, P. R., A. Lunde, and J. M. Nason (2011): “The model confidence set,” *Econometrica*, 79(2), 453–497.

Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman (2009): *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer.

Husted, L., J. Rogers, and B. Sun (2020): “Monetary policy uncertainty,” *Journal of Monetary Economics*, 115, 20–36.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013): *An introduction to statistical learning*, vol. 112. Springer.

Jurado, K., S. C. Ludvigson, and S. Ng (2015): “Measuring uncertainty,” *American Economic Review*, 105(3), 1177–1216.

Li, J., Z. Liao, and R. Quaedvlieg (2022): “Conditional superior predictive ability,” *The Review of Economic Studies*, 89(2), 843–875.

Mullainathan, S., and J. Spiess (2017): “Machine learning: an applied econometric approach,” *Journal of Economic Perspectives*, 31(2), 87–106.

Phillips, P. C., and P. Perron (1988): “Testing for a unit root in time series regression,” *Biometrika*, 75(2), 335–346.

Welch, I., and A. Goyal (2007): “A comprehensive look at the empirical performance of equity premium prediction,” *The Review of Financial Studies*, 21(4), 1455–1508.