Empirical investigation on supervised machine learning models predicting equity risk premium

Myong Jong Shin*

November 28, 2022

Abstract

We examine the predictive performance of supervised machine learning models in forecasting multi-horizon firm-level equity risk premium. A large collection of individual firms' financial characteristics and US macroeconomic predictors for returns from January 1960 to December 2019 are used. We forecast excess returns for (1) all individual firms and for (2) each group of firms belonging to the same industry sector in US. We first show out-of-sample fit for each forecast model. Second, we forecast evaluate models to find ones with superior predictive ability and are included in model confidence sets. Finally we test for conditional superior predictive ability of a model, where a model's CSPA is conditional on a prior chosen variable indicative of the state of the market.

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 research topic. In this article, we conduct a comparative analysis of supervised machine learning models forecasting multi-horizon equity risk premium for individual firms in the US stock market. The goal is to identify a model or a set of models that have better return predictability over others with statistical significance. We do so by employing a set of out-of-sample forecast evaluation tests from highly influential papers in recent years. We study both for the entire set of stocks collected from the US stock markets, and for groups of stocks within our original set that are categorized with respect to their industry classification.

Predicting excess returns is a difficult task. The field has seen a plethora of forecasting methods, a combination of forecasting models and predictors used, reported to have good return predictability, but they are often sensitive to the choice of time periods of the sample and the ranking of different forecasting methods for their predictability is time varying. For example, Figure 1 shows the squared forecast error from predicting monthly excess stock returns with a naive forecast of zeros, henceforth called 0forecast, from January 1987 to December 2019. To forecast a firm i's excess returns, the prediction of 0forecast is $\hat{r}_{it} = 0, \forall t$.

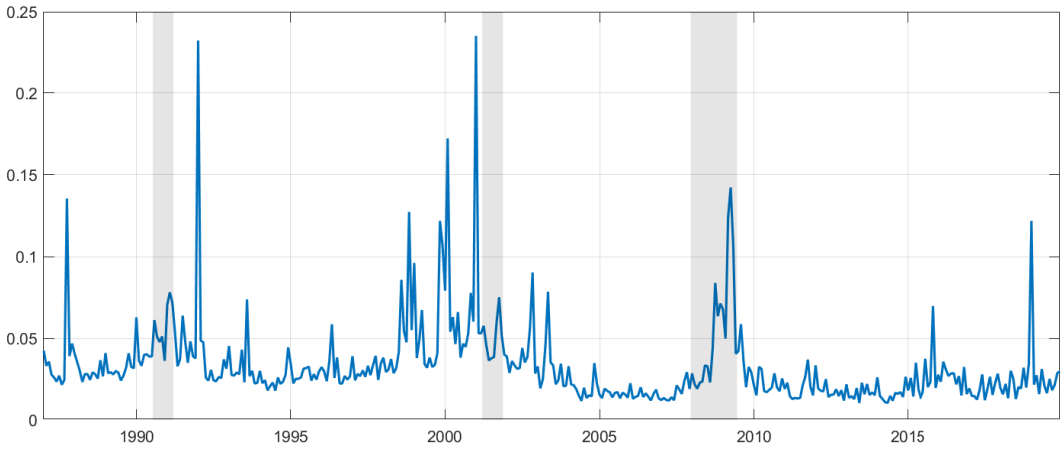


Figure 1: Squared equity risk premium of US stock market.

It is clear that forecast errors are large over certain historical events in the 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. We can expect that under certain time periods there will be models that can better forecast excess returns.

Given the noisy nature of the financial markets, we do not expect to find a single model or a fixed set of models that exhibit consistent return predictability across time and robust to different choices of firms or financial market. Thus, a performance evaluation of the forecasting methods in a frequent basis is important to find our bearings and adapt our tools for future research in forecasting excess returns.

In their influential paper, Gu, Kelly, and Xiu (2020), henceforth GKX, surveys a set of supervised machine learning models to forecast monthly equity risk premium for individual firms in the US stock market. Their data is extensive in time ranging from March 1957 to December 2016, and in the number of firms collected with the average number of stocks per month exceeding 6,200. We are highly motivated by their research and provide our contributions in the following ways.

First, we extend their analysis to forecast different multi-horizon returns. In this article we study monthly, quarterly, semi-annual, and annual returns. The list of good models for predictability is different for each returns. Second, we conduct forecast evaluation tests to determine superior predictive ability(SPA) and conditional superior predictive ability(CSPA) for a model and construct model confidence sets. Finally, we check return predictability for different industry sectors in the US economy and find models with SPA and CSPA. We find that for some industries, the set of models with good predictability is different than those for the entire market.

The rest of the paper is organized as follows. Section 2 the supervised machine learning models used in our article. Section 3 describes the out-of-sample forecast evaluation tests for SPA and CSPA. Section 4 describes the data used. Section 5 shows the results and Section 6 concludes.

2 Supervised Machine Learning 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 1. 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 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.

Individual firm's h-horizon excess return is calculated as continuously compounded stock returns minus the risk free rate written as $\sum_{\tau=0}^{h-1} r_{i,t+\tau}$ where h indicates the multi-horizon dimension in number of months. In our study we forecast for one month ahead ($h = 1$), one quarter ahead ($h = 3$), six month ahead ($h = 6$), and one year ahead ($h = 12$) excess returns. The firms are indexed by $i = 1, \cdots N_t$ where the total number of firms changes and indexed by $t = 1, \cdots T$. All models use a common matrix of covariates $z_{i,t-1}$. θ is a coefficient vector of size P. Let $\mathbb{J}$ be the set of supervised machine models used in this article. All our models are additive prediction error regression model for firm i's return written in Equation 1. We forecast returns by the estimated conditional expectation of the regression.

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

We denote the different supervised machine learning models by subscript $j \in \mathbb{J}$. A firm i's h-horizon excess returns is predicted as $f_j(z_{i,t-1}, \hat{\theta})$ for model j as the minimizer of an objective function such as least squares. For example, when using partial least squares for prediction, a model in our set of supervised machine learning models, we denote its forecast as $f_{pls}(z_{i,t-1}, \hat{\theta})$. The 0forecast model for forecasting multi-horizon excess returns is $\sum_{\tau=0}^{h-1} r_{i,t+\tau} = 0$ requiring no estimation, and denoted with f_0.

For our out-of-sample scheme, we split our full sample 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 forecast errors.

2.1 Principal component regression

Principal component regression(PCR) uses principal component(PC) ω_k. PCs are extracted from the sample covariance matrix of predictors Z. The solution to optimization problem in Equation 2 is the eigenvector associated with kth largest eigenvalue.

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

For each estimation using PCs from 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 number of PCs and the chosen number of components is used for prediction during the testing period. 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 that PCs explaining 60% or lower level of the variance of covariates provide better predictability.

2.2 Partial least squares regression

Partial least squares regression(PLS) extracts PLS components ω_k by solving Equation 3. Whereas PCR extracts linear combinations of predictors that best explain the covariance matrix of predictors, PLS extracts linear combinations of predictors that best explain the squared covariance between the multi-horizon excess returns and the matrix of covariates Z.

$$\omega_k = \arg\max_{\omega} Cov^2(\sum_{\tau=0}^{h-1} r_{t+\tau}, Z\omega), \;\; \omega'\omega = 1, \;\; Cov(Z\omega, Z\omega_k) = 0, \;\; k = 1, 2, \cdots, P-1. \tag{3}$$

To solve Equation 3 we use SIMPLS algorithm by De Jong (1993). We use the validation data and choose the number of PLS components with the smallest mean squared error during the validation period. The chosen number of components is then used for prediction for the testing period.

For both PCR and PLS, least squares objective function is used to estimate the coefficient vector θ

and we make prediction $f(z_{i,t-1}, \hat{\theta})$. Equation 4 shows the resulting pooled least squares estimator.

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{N_t} \sum_{t=1}^{T} \Big(\sum_{\tau=0}^{h-1} r_{i,t+\tau} - f_j(z_{i,t-1}; \theta) \Big)^2 \tag{4}$$

2.3 LASSO

LASSO stands for least absolute shrinkage and selection operator. The model includes a penalty function in addition to the least squares objective function to find the unique optimizer for coefficient vector θ. There are many variants of the LASSO with different penalty functions. Tibshirani (2011) provides an excellent review of the different generalizations of the LASSO method. For our article, we use the most simple absolute penalty and our LASSO problem can be stated as Equation 5 in lagrangian form with $\lambda \geq 0$.

$$\hat{\theta}_{lasso} = \arg\min_{\theta} \Big(\sum_{i=1}^{N_t} \sum_{t=1}^{T} \Big(\sum_{\tau=0}^{h-1} r_{i,t+\tau} - f_j(z_{i,t-1}; \theta) \Big)^2 + \lambda \sum_{k=1}^{p} |\theta_k| \Big) \tag{5}$$

By constraining θ by $l1$ norm, LASSO induces *sparsity* where we can 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.

2.4 Ridge

Ridge model includes a squared penalty function to the least squared objective function that constrains the magnitude of θ and its problem can be stated as Equation 6 in lagrangian form with $\lambda \geq 0$.

$$\hat{\theta}_{ridge} = \arg\min_{\theta} \Big(\sum_{i=1}^{N_t} \sum_{t=1}^{T} \Big(\sum_{\tau=0}^{h-1} r_{i,t+\tau} - f_j(z_{i,t-1}; \theta) \Big)^2 + \frac{1}{2} \lambda \sum_{k=1}^{p} \theta_k^2 \Big) \tag{6}$$

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.

2.5 Elastic Net

The elastic net model from Zou and Hastie (2005) 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.

$$\hat{\theta}_{enet} = \arg\min_{\theta} \Big(\sum_{i=1}^{N_t} \sum_{t=1}^{T} \big(\sum_{\tau=0}^{h-1} r_{i,t+\tau} - f_j(z_{i,t-1}; \theta) \big)^2 + \lambda(1-\rho) \sum_{k=1}^{p} |\theta_k| + \frac{1}{2} \lambda \rho \sum_{k=1}^{p} \theta_k^2 \Big) \tag{7}$$

2.6 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. There are many ways to grow the random forest depending on the prediction exercise at hand. For our article, 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. A branch separates the data from the previous branch using one of the predictor variables. Its goal is to group observations that are similar to each other in to binary bins. At branch C, a tree chooses a predictor in $z^b_{i,t}$ that can split the data that minimize $l2$ impurity in Equation 8 where $|C|$ denotes the number of observations at branch C.

$$H(\theta, C) = \frac{1}{|C|} \sum_{z^b_{i,t-1} \in C} (\sum_{\tau=0}^{h-1} r^b_{i,t+\tau} - \theta)^2 \tag{8}$$

We stop splitting branches 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 $f_{tree,b}(z^b_{i,t-1}, \hat{\theta}^b, L)$. The final random forest prediction bags predictions from individual trees as their average in Equation 10.

$$f_{tree,b}(z^b_{i,t-1}, \hat{\theta}^b, L) = \sum_{k=1}^{2^L} \hat{\theta}^b_k \mathbb{1}\{z^b_{i,t-1} \in C_k(L)\}, \quad \hat{\theta}^b_k = \frac{1}{|C_k(L)|} \sum_{z^b_{i,t-1} \in C_k(L)} \sum_{\tau=0}^{h-1} r^b_{i,t+\tau} \tag{9}$$

$$f_{rf} = \frac{1}{B} \sum_{b=1}^{B} f_{tree,b}(z^b_{i,t-1}, \hat{\theta}^b, L) \tag{10}$$

3 Forecast evaluation

Here we discuss the statistical tests used to evaluate performance of forecast models. A common statistic used in machine learning literature to evaluate forecasts from a model is an out-of-sample fit. It is convenient to calculate and is widely used by financial economists, statisticians, and researchers in computer science when analyzing predictive performance of a model. However, we argue that with the tests used in our article, researchers can make better comparisons across models and avoid multiple hypothesis testing problem. Moreover, test for CSPA introduces a new way to evaluate forecasts conditional on the state of the economy.

3.1 R^2_{OOS}

R^2_{OOS} statistic is the out-of-sample fit during the testing sample period $\mathbb{T}_3$ for each model. First we pool forecast errors across time and firms to calculate the mean squared forecast error from model j and the 0forecast. Then we measure the reduction in mean squared forecast error for using model j relative to 0forecast. Outperforming it on average is shown by a positive R^2_{OOS} value and a negative value indicates the relative underperformance of model j.

$$R^2_{\text{OOS},j} = 1 - \frac{\sum_{(i,t)\in\mathbb{T}_3} \left(\sum_{\tau=0}^{h-1} r_{i,t+\tau} - f_j(z_{i,t-1}, \hat{\theta})\right)^2}{\sum_{(i,t)\in\mathbb{T}_3} \left(\sum_{\tau=0}^{h-1} r_{i,t+\tau}\right)^2} \tag{11}$$

We use the definition of R^2_{OOS} from Gu, Kelly, and Xiu (2020). This formula is different from the conventional one for out-of-sample fit that uses historical mean forecast rather than the 0forecast. The reasons for this choice are twofold. First, this formula for R^2_{OOS} provides information regarding a model's out-of-sample fit relative to a model used in our study. Thus the interpretation of the statistic is more meaningful. Second, outperforming historical mean forecast is typically a less challenging task for the models studied here, due to the noisy nature of the firm-level return data.

3.2 Test for Superior Predictive Ability and Model Confidence Set

We use squared forecast error to measure a prediction's accuracy for all our models[1]. Given h at time t, we have N_t number of errors per model. Let $\mathbb{L}_{j,t+h-1}$ be the cross sectional average of squared error from model j for all firms.

[1] 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 thus penalizing them less.

$$\mathbb{L}_{j,t+h-1} = \frac{1}{N_t}\sum_{i=1}^{N_t}\left[\left(f_j(z_{i,t-1},\hat{\theta}) - \sum_{\tau=0}^{h-1} r_{i,t+\tau}\right)^2\right] \tag{12}$$

We say that the model j has superior predictive ability (SPA) over model j' when the squared forecast error of model j is smaller than model j' on average. We denote $\mathbb{L}_{j,t+h-1} - \mathbb{L}_{j',t+h-1}$ as $\Delta\mathbb{L}_{j,j',t+h-1}$ for simplicity and use them interchangeably.

$$\mathbb{E}(\Delta\mathbb{L}_{j,j',t+h-1}) \leq 0 \tag{13}$$

For example, Figure 2 plots the difference between partial least square model's squared forecast error and that of the 0forecast for monthly forecasts, $\mathbb{L}_{pls,t} - \mathbb{L}_{0forecast,t}$. During time periods where $\Delta\mathbb{L}_{pls,0forecast,t} < 0$ on Figure 2, the partial least squares has smaller squared error thus outperforming 0forecast. If $\mathbb{E}\Delta\mathbb{L}_{pls,0forecast,t} < 0$, we say that partial least squares has SPA over 0forecast.

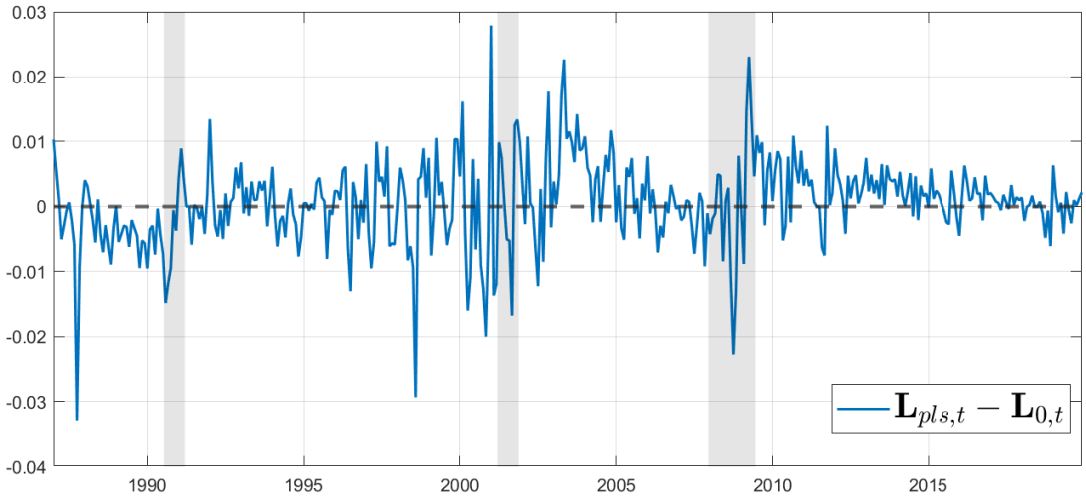


Figure 2: $\Delta\mathbb{L}_{pls,0forecast,t}$, January 1987 to December 2019.

Prior to our tests, for all pair of models, we checked for existence of unit-root for the time series $\Delta\mathbb{L}_{j,j',t+h-1}$ for all pairs of models, as all the tests in this article assumes stationary. This does not mean that the data and forecasts need to be stationary. Also, as all our models are additive prediction error regressions in the form of Equation 1, the error terms are additive. We take into account the serial correlation of the errors for all our tests using NeweyWest standard errors with lag $h-1$.

The main research questions regarding SPA are twofold. First, for a pair of models $(j, j') \in \mathbb{J}$, we test significance for all model pairs under null of $\mathbb{E}(\Delta\mathbb{L}_{j,j',t+h-1}) \leq 0$. We test this null using test of SPA by Hansen (2005). Second, we construct a model confidence set(MCS) from Hansen, Lunde, and Nason (2011) that contains models with SPA over all other models with a chosen coverage probability.

To test the null of superior predictive ability(SPA) of model j over j' for all pairwise combinations of models, we use the test for SPA by Hansen (2005). For the difference of mean squared forecast error between model j and j' during testing period $\bar{d}_{j,j',t+h-1} = 1/\mathbb{T}_3 \sum \Delta\mathbb{L}_{j,j',t+h-1}$, we describe the testing procedure.

1. Studentize the test statistic with NeweyWest standard error with lag $h-1$ and normalized to 0.

$$T_{SPA} = \max\left[\frac{\sqrt{\mathbb{T}_3}\bar{d}_{j,j',t+h-1}}{\sqrt{var(\sqrt{\mathbb{T}_3}\bar{d}_{j,j',t+h-1})}}, 0\right]$$

2. Bootstrap the studentized test statistic but recenter it to $\hat{\mu}^c$ where

$$\begin{cases} \hat{\mu}^c = \bar{d}_{j,j',t+h-1}, \;\; if \frac{\sqrt{\mathbb{T}_3}\bar{d}_{j,j',t+h-1}}{\sqrt{var(\sqrt{\mathbb{T}_3}\bar{d}_{j,j',t+h-1})}} \leq -\sqrt{2\log\log\mathbb{T}_3} \\ \hat{\mu}^c = 0, \;\; otherwise \end{cases}$$

When model j has the better sample performance, the test statistic is normalized to 0 and we conclude we find no evidence against the null hypothesis, and consequently the null of $H_0 : \mathbb{E}(\Delta\mathbb{L}_{j,j',t+h-1}) \leq 0$ should not be rejected. The resampled test statistic is recentered using the law of iterated logarithms to generate the bootstrap distribution that conforms with the null hypothesis. For implementation, we use the stationary bootstrap of Politis and Romano (1994).

Next, we construct a model confidence set(MCS) from Hansen, Lunde, and Nason (2011) with 5% significance level to collect a set of models with SPA. With MCS, we can conduct *all-for-one comparison* of models, where we search for models with better forecast performance over all other models on average jointly. To define MCS, we assume that $\mu_{j,j',h-1}$, the unconditional expectation of the difference of mean squared forecast error between model j and j' exists and finite.

$$\mu_{j,j',h-1} = \mathbb{E}\Delta\mathbb{L}_{j,j',t+h-1}, \quad \forall j, j' \in \mathbb{J} \tag{14}$$

For all model pairs, the set of superior models $\mathbb{M}^*$ is

$$\mathbb{M}^* = \{j \in \mathbb{J} : \mu_{j,j',h-1} \leq 0, \forall j' \in \mathbb{J}\} \tag{15}$$

Using the MCS algorithm in Hansen, Lunde, and Nason (2011) we construct the set of superior models and we use the stationary bootstrap for p-values.

3.3 Conditional Superior Predictive Ability

We test for CSPA of a model using the testing procedure from Li, Liao, and Quaedvlieg (2022). Testing for CSPA of a model means that we evaluate the forecast model with another for their relative predictive ability conditional on some chosen state variable. An unconditional comparison of average performance we made for testing SPA and constructing MCS can be considered as a special case of testing for CSPA, given a pair of models or a group of multiple models. Because we only compare the average relative performance of models for SPA and MCS, we integrated out a possible heterogeneity across important periods during the testing sample in the data such as recession periods. The purpose of CSPA is to discover the state-dependent relative performance among forecast models.

The null hypothesis of CSPA in our article is that the conditional expected squared error of a model j is no larger than those of all other competing models uniformly across all conditioning states, thus outperforming all competitions uniformly[2]. Formally, given a forecast model $j \in \mathbb{J}$, we test the null of CSPA of the model j over all $j' \in \mathbb{J}$, given a priori chosen state variable X_t. Note that by the law of iterated expectations the null of CSPA implies the null of model j's SPA over all others.

$$H_0 : \mathbb{E}(\Delta\mathbb{L}_{j,j',t+h-1}|X_t = x_t) \leq 0, \ \ \forall j' \in \mathbb{J}, X_t \in \mathbb{X} \tag{16}$$

$\mathbb{X}$ is the support of the state variable. The conditioning state must be a scalar for implementing the test. Table 1 lists the different state variables collected and used in our article with and their sources in parenthesis. We report CSPA results using all the state variables for robustness. For our article we use business cycle indicators, financial conditions or stress indicators, and uncertainty indices.

[2]Unlike the original null hypothesis of CSPA stated in Li, Liao, and Quaedvlieg (2022), the direction of the inequality sign in our article is switched to be coherent with our analysis of testing SPA and constructing MCS.

Table 1: List of state variables

Mnemonic	
ADS	Aruoba-Diebold-Scotti index(Philadelphia Fed)
ANFCI	Adjusted national financial conditions index(Chicago Fed)
NFCI	National financial conditions index(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	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))

Our purpose of using CSPA to find evidence of an uniform conditional dominance of a forecast model. We consider that a model having CSPA over all others for different state variables to be a strong indication that it is the model with the superior predictive performance robust to different states of the economy.

To implement the test, the conditional expectation function $\mathbb{E}(\Delta\mathbb{L}_{j,j',t+h-1}|X_t = x_t)$ is approximated via a nonparametric series regression of $\Delta\mathbb{L}_{j,j',t+h-1}$ on the basis expansion of the state variable X_t using legendre polynomials. To calculate the correct critical values for the test, Li, Liao, and Quaedvlieg (2022) provides an algorithm for implementation. We take serial correlation into account for the test by using NeweyWest standard errors.

4 Data

To forecast multi-horizon excess returns we use a large collection of firm characteristics and macroeconomic predictors for the aggregate market return. The key paper that serves as a main motivation for data construction in our article is GKX. GKX use 94 firm characteristics from Green, Hand, and Zhang (2017), 8 macroeconomic predictors from Welch and Goyal (2007), and 74 dummy variables for classification of firms to different industries define by the first two digits of the US Standard Industrial Classification (SIC) Codes. With the data from March 1957 to December 2016, GKX predict individual monthly excess returns using supervised machine learning models check each model's fit out-of-sample. For our analysis, we use the data from GKX with more recent observations of predictors included up to December 2019[3].

[3]The data for predictors in our article is provided by the authors and is readily available at their website but requires access to high performance computation and large storage to handle the big data and heavy computation burden.

Our 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 SPA and CSPA. Moreover we forecast multi-horizon excess returns by firm groups categorized by their industry sector. To do so, we do not include SIC dummies from GKX as predictors. We instead forecast excess returns per industry using firm SIC codes and compare forecast model within. We use the first digit of each firm's SIC code and categorize the industries in US into 10 different sectors. Table 2 lists the class of firms by their first digit of the SIC code into 10 industry sectors. We use the same set of forecast models from studying the entire market to each industry and report the SPA and CSPA test results.

Table 2: Standard Industrial Classification Codes

First digit of SIC	Description
0	Agriculture, Forestry and Fishing
1	Mining and Construction
2,3	Manufacturing
4	Transportation, Communications, Electric, Gas and Sanitary service
5	Wholesale Trade and Retail Trade
6	Finance, Insurance and Real Estate
7,8	Services
9	Public Administration and Nonclassifiable

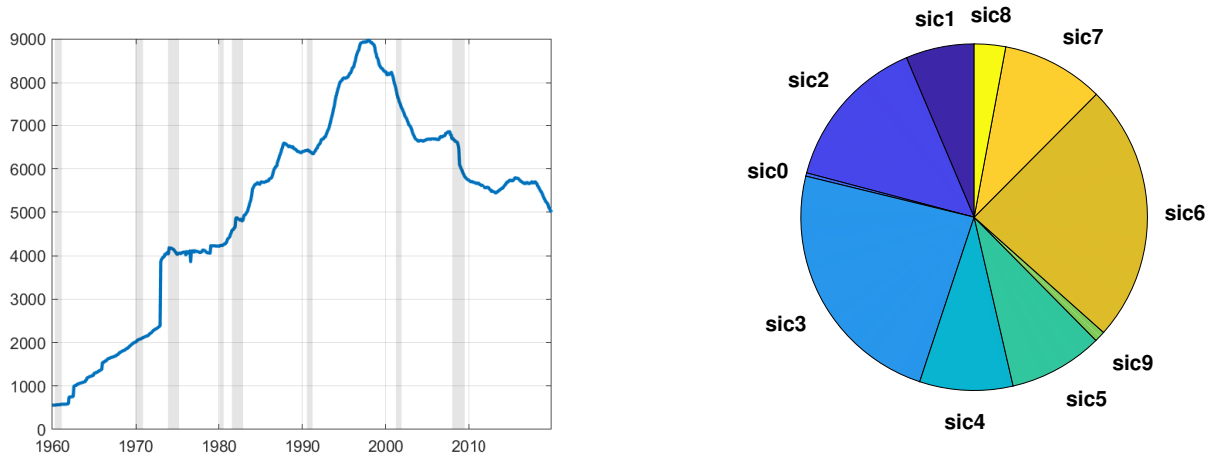
Let s denote a set of firms in the same industry sector. For example, for a set of firms belonging to *Real Estate industry*, they can be labeled with $s = 6$. Given h at time t, we have $N_{t,s}$ number of firms per industry. For forecast model j, we can define the cross sectional average of squared error for industry s as $\mathbb{L}_{s,j,t+h-1}$. We use this to test for SPA and CSPA of models within an industry.

$$\mathbb{L}_{s,j,t+h-1} = \frac{1}{N_{t,s}} \sum_{i \in s}^{N_{t,s}} \left[\left(f_j(z_{i,t-1}, \hat{\theta}) - \sum_{\tau=0}^{h-1} r_{i,t+\tau} \right)^2 \right] \tag{17}$$

$$s = 0, 1, \cdots, 9 \quad \sum_{s=0}^{9} N_{t,s} = N_t \tag{18}$$

Monthly firm equity returns are from CRSP for all firms listed in NYSE, AMEX, and NASDAQ. We collect our data from January 1960 to December 2019. Each firm's excess returns are calculated as individual continuously compounded stock returns minus the risk free rate, and we use treasury-bill rate to proxy the 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. The proportion of each industry sector

in our sample is shown alongside as a pie chart. The proportion for each industry sector s is calculated as the average across time and cross section $\frac{1}{T}\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{1}\big(\text{firm } i \in \text{Industry sector } s\big)$. The largest proportion of industry in our data is *Manufacturing* with SIC code 2 and 3, and the smallest is *Agriculture, Forestry and Fishing* with SIC code 0.



(a) Number of firms N_t.

(b) Proportion of industries.

Figure 3: Summary of all stocks from Jan1960 to Dec2019

For predictors, we have two groups of covariates. First, we have 94 firm level characteristics from GKX. Our samples contains over 3 million pooled observations for each firm characteristic[4]. For firm i at time t, let $c_{i,t}$ be a 94 by 1 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) for market equity premium and they consist of market stock characteristics and bond interest-related predictors. For our research we use dividend-price ratio, earning-price ratio, book-to-market ratio, net equity expansion, stock variance, treasury-bill rate, term spread, and default spread. At time t, let m_t is a 8 by 1 vector of macroeconomic predictors.

With $c_{i,t}$ and m_t, we construct the vector of predictors $z_{i,t}$ use for all supervised machine learning models. Let P be the total number of covariates and $z_{i,t}$ be a P by 1 vector. Then,

$$z_{i,t} = \big[1,\ m_t\big] \otimes c_{i,t} \tag{19}$$

Therefore $P = 94 * (8 + 1) = 846$ including the interaction between $c_{i,t}$ and m_t. We use $z_{i,t}$ to make

[4]Each firm characteristic is winsorized cross sectionally for 1% and 99%. Refer to the Appendix of Gu, Kelly, and Xiu (2020) for full description of the firm characteristics used.

predictions for model j using Equation 1.

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

For our out-of-sample analysis, we use the rolling scheme. At the beginning, we have the initial 15 years of training period from January 1960 to December 1974, and 12 years of validation from January 1975 to December 1986. We then forecast excess 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 15 years of training period from January 1961 to December 1975, and a new 12 years of validation from January 1976 to December 1987. We then make our forecast for the entire year of 1988. We repeat this process until the end of the sample. In the end we have 33 years of testing period from January 1987 to December 2019 that we can compare with observed excess returns and calculate forecast errors.

5 Results

First, we forecast multi-horizon returns for all firms in our data using supervised machine models described in Section 2. For returns of horizon h, $h = 1$ is monthly, $h = 3$ is quarterly, $h = 6$ is semi-annual, and $h = 12$ is annual excess returns.

Table 3: All firms: R^2_{oos}, MCS results

	h=1		h=3		h=6		h=12	
	r2	MCS	r2	MCS	r2	MCS	r2	MCS
0forecast	-	✓	-		-		-	
Enet	-0.0176		0.0140		0.0214		0.0257	
LASSO	-0.0177		0.0144		0.0217		0.0272	
PCR	-0.0176		0.0020		0.0056		0.0061	
PLS	-0.0171		0.0180		0.0260		0.0324	
Ridge	-0.0648		-0.0590		-0.0942		-0.0837	
RF	-0.0092		0.0351	✓	0.0377	✓	0.0397	✓

Models in MCS with 5% marked with ✓, stationary bootstrap with B=100000 and block length=5000.

Table 3 shows R^2_{OOS} of each model and the models included in MCS with 5% significance. We observe that as the number of horizon h increases, R^2_{OOS} increases for most models. Aside from predicting monthly excess returns our result shows that the random forest model shows better out-of-sample fit relative to others and is the only one included in the MCS. This suggests that when forecasting multi-horizon returns for all the firms collected in our data, random forest should be chosen among the models tests in our article.

The ridge model show poor out-of-sample fit for all horizons and we suspect this is due to the sheer

number of covariates used that shrinkage alone could not handle. This is further supported by observing that for elastic net model and LASSO model, there fit is very similar suggesting that elastic net mostly benefited from variable selection, reducing the number of parameters to estimate.

Table 4: All firms: SPA results(pairwise)

		0forecast	Enet	LASSO	PCR	PLS	Ridge	RF
	0forecast	-						
	Enet	-	-					RF
	LASSO	-	-	-				RF
h=1	PCR	-	-	-	-			RF
	PLS	-	-	-	-	-		RF
	Ridge	-	-	-	-	-	-	RF
	0forecast	-	Enet	LASSO		PLS		RF
	Enet	-	-	LASSO		PLS		RF
	LASSO	-	-	-		PLS		RF
h=3	PCR	-	-	-	-	PLS		RF
	PLS	-	-	-	-	-		RF
	Ridge	-	-	-	-	-	-	RF
	0forecast	-	Enet	LASSO		PLS		RF
	Enet	-	-			PLS		RF
	LASSO	-	-	-		PLS		RF
h=6	PCR	-	-	-	-	PLS		RF
	PLS	-	-	-	-	-		RF
	Ridge	-	-	-	-	-	-	RF
	0forecast	-	Enet	LASSO		PLS		RF
	Enet	-	-			PLS		RF
	LASSO	-	-	-		PLS		RF
h=12	PCR	-	-	-	-	PLS		RF
	PLS	-	-	-	-	-		
	Ridge	-	-	-	-	-	-	RF

For a model j(in row) that has SPA over model j'(in column), its name is written.

Table 4 displays test results for SPA between model pairs with 5% significance. For readability, we write the name of the model that has SPA over the other in the table. For example, at $h = 3$, between 0forecast and elastic net model, the latter has SPA. We can see that random forest has SPA over almost every model. Dimension reduction technique such as partial least squares and model selection via elastic net model or LASSO model also has SPA over 0forecast and some other models.

Next, we report test results for a model's CSPA in Table 5. A ($\star$) indicates rejection of the null hypothesis of a model j's CSPA over all other competing models uniformly given a conditioning state variable; $\mathbb{E}(\Delta\mathbb{L}_{j,j',t+h-1}|X_t = x_t) \leq 0,\ \forall j' \in \mathbb{J}, X_t \in \mathbb{X}$. Names of the model j being tested is written on the left-column, and the names of each state variable used per a test is written on the top-row. For example, in $h = 12$, the random forest model has CSPA over all other forecast models uniformly for all our state variables. Uncertainty indices from Jurado, Ludvigson, and Ng (2015) for semi-annual returns are not available and thus not reported.

Table 5: All firms: CSPA results

		ADS	ANFCI	NFCI	rec prob	vxo	gecon	mpu	epu	JLN f	JLN r	JLN m
	0forecast											
	Enet	⋆	⋆	⋆		⋆	⋆	⋆	⋆	⋆	⋆	⋆
	LASSO	⋆	⋆	⋆		⋆	⋆	⋆	⋆	⋆	⋆	⋆
h=1	PCR	⋆	⋆	⋆		⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS	⋆	⋆	⋆		⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF		⋆			⋆						
	0forecast	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Enet	⋆	⋆	⋆		⋆	⋆	⋆	⋆		⋆	⋆
	LASSO										⋆	⋆
h=3	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS										⋆	
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF											
	0forecast	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	NA	NA	NA
	Enet									NA	NA	NA
	LASSO									NA	NA	NA
h=6	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	NA	NA	NA
	PLS									NA	NA	NA
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	NA	NA	NA
	RF									NA	NA	NA
	0forecast			⋆		⋆		⋆	⋆		⋆	
	Enet	⋆	⋆	⋆	⋆	⋆	⋆			⋆	⋆	⋆
	LASSO	⋆	⋆	⋆	⋆		⋆			⋆	⋆	⋆
h=12	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS									⋆		
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF											

A model that reject CSPA null with 5% significance marked (⋆).

We can see that the random forest model generally has CSPA over all others in almost all horizons and state variables. This complements the results from before and provide more evidence that random forest model is a good choice of model in our data. Partial least squares model exhibit CSPA except for $h = 1$. Elastic net model and LASSO has CSPA in $h = 6$, and principal component regression model and ridge model is rejected the most often suggesting that they are often outperformed by others under certain states in the testing period.

Next, we repeat our forecast evaluation exercise for predicting firms belonging to the same industry sectors in the US. We define different industry sectors by a firm's first digit of the SIC code, and we have a total to 10 sectors. We do not report pairwise SPA test results due to difficulty in readability of results.

Table 6 reports out-of-sample fit and MCS for each industry sector and horizon. The industries are labeled on the left column as sic#. For example sic0 means the models forecasts for firms whose first digit of the SIC code is 0, the *Agriculture, Forestry and Fishing* sector. We will refer each industry by

Table 6: By industry: R^2_{oos}, MCS results

		h=1		h=3		h=6		h=12	
		R^2_{oos}	MCS	R^2_{oos}	MCS	R^2_{oos}	MCS	R^2_{oos}	MCS
	0forecast	-	✓	-		-		-	
	Enet	-0.0095		0.0024		0.0161		0.0267	
	LASSO	-0.0090	✓	0.0023		0.0149		0.0257	
sic0	PCR	-0.0201		0.0072		0.0119		0.0040	
	PLS	-0.0152		0.0152	✓	0.0230	✓	0.0611	✓
	Ridge	-0.0202		-0.0092		-0.0086		-0.0063	
	RF	-0.0080	✓	0.0126	✓	0		0.0288	
	0forecast	-	✓	-		-		-	
	Enet	-0.0211		0.0191	✓	0.0142	✓	0.0069	
	LASSO	-0.0212		0.0191	✓	0.0139		0.0084	
sic1	PCR	-0.0242		0.0030		0.0081		0	
	PLS	-0.0209		0.0056		0.0081		0.0128	✓
	Ridge	-0.0326		-0.0369		-0.0417		-0.0691	
	RF	0.0153	✓	0.0424	✓	0.0094		-0.0268	
	0forecast	-		-		-		-	
	Enet	0.0090	✓	0.0086		0.0069		0.0074	
	LASSO	0.0089		0.0086		0.0065		0.0099	
sic2	PCR	0.0051		0.0023		0.0053		0.0040	
	PLS	0.0088		0.0098		0.0025		0.0016	
	Ridge	-0.0140		-0.0372		-0.0698		-0.0877	
	RF	0.0031		0.0753	✓	0.0580	✓	0.0654	✓
	0forecast	-		-		-		-	
	Enet	0		0.0175		0.0201		0.0269	
	LASSO	0		0.0175		0.0199		0.0281	
sic3	PCR	-0.0084		0.0061		0.0112		0.0138	
	PLS	0		0.0166		0.0207		0.0295	
	Ridge	-0.0374		-0.0658		-0.1077		-0.0999	
	RF	0.0757	✓	0.0597	✓	0.0620	✓	0.0836	✓
	0forecast	-		-		-		-	
	Enet	0.0095		0.0115		0		-0.0012	
	LASSO	0.0094		0.0115		0		0.0035	
sic4	PCR	0.0070		0.0048		0.0109		0.0093	
	PLS	0.0086		0.0077		-0.0086		-0.0139	
	Ridge	-0.0275		-0.0554		-0.0952		-0.1158	
	RF	0.3095	✓	0.3034	✓	0.2855	✓	0.2514	✓
	0forecast	-		-		-		-	
	Enet	0.0154		0.0112		-0.0033		-0.0023	
	LASSO	0.0153		0.0111		-0.0038		0.0032	
sic5	PCR	0.0130		0.0066		0.0139		0.0141	
	PLS	0.0133		0.0092		-0.0136		-0.0204	
	Ridge	-0.0305		-0.0604		-0.1033		-0.1205	
	RF	0.1526	✓	0.1504	✓	0.1582	✓	0.1892	✓
	0forecast	-		-		-		-	
	Enet	0.0115		0.0138		0.0138		0.0202	
	LASSO	0.0114		0.0137		0.0132		0.0197	
sic6	PCR	0.0041		0.0053		0.0092		0.0084	
	PLS	0.0106		0.0148		0.0162		0.0210	
	Ridge	-0.0225		-0.0565		-0.1090		-0.1401	
	RF	0.6098	✓	0.6210	✓	0.5961	✓	0.5918	✓
	0forecast	-		-		-		-	
	Enet	0.0151		0.0101		0.0054		0.0075	
	LASSO	0.0150		0.0101		0.0052		0.0108	✓
sic7	PCR	0.0132		0.0050		0.0095	✓	0.0101	✓
	PLS	0.0160	✓	0.0124	✓	0.0036		0.0028	
	Ridge	-0.0133		-0.0520		-0.0968		-0.1175	
	RF	-0.2658		-0.2571		-0.2503		-0.2576	
	0forecast	-		-		-		-	
	Enet	0.0137		0.0225		0.0344		0.0498	
	LASSO	0.0137		0.0226		0.0347		0.0501	✓
sic8	PCR	0.0183		0.0140		0.0251		0.0333	
	PLS	0.0218		0.0245		0.0416	✓	0.0520	✓
	Ridge	-0.0142		-0.0508		-0.0515		-0.0412	
	RF	0.1521	✓	0.0654	✓	0.0544	✓	0.0544	✓
	0forecast	-		-		-		-	
	Enet	0.0472		0.0100		-0.0043		0	
	LASSO	0.0472		0.0100	✓	-0.0046		-0.0012	
sic9	PCR	0.0608		0.0081		0.0114	✓	0.0096	✓
	PLS	0.0650	✓	0.0105	✓	0.0097		0.0037	
	Ridge	0.0228		-0.0856		-0.1211		-0.1693	
	RF	-1.0684		-0.1537		-0.1451		-0.0889	

Models included in MCS with 5% significance marked with ✓.

the mnemonic used in Table 6 for simplicity.

First, we observe that for industries sic3 to sic 6, the random forest model has good out-of-sample fit and the sole member of MCS for all horizons. However for industries sic7 and sic9, it has worse fit than others and never included in MCS. This change in the model with the best fit could not have been observed if we had not divided our firms into different groups. Concurrently, the group of firms whose first digit SIC code is 3, 4, 5, or 6 makes up the majority of the firms as it can be seen in proportions from Figure 3.

Next, we report test results for a model's CSPA by industry. A ($\star$) indicates rejection of the null hypothesis of a model j's CSPA over all other competing models uniformly given a conditioning state variable; $\mathbb{E}(\Delta\mathbb{L}_{j,j',t+h-1}|X_t = x_t) \leq 0,\ \forall j' \in \mathbb{J}, X_t \in \mathbb{X}$. Names of the model j being tested is written on the left-column, and the names of each state variable used per a test is written on the top-row. For example, in Table 7 for $h = 1$ and sic0, the LASSO has CSPA over all other forecast models uniformly for all our state variables.

Table 7, 8, 9, and 10 shows the CSPA results of models for each industry and returns. We see that there is no single model or a fixed group of models that has CSPA across all industries and horizons. This provides evidence of heterogeneity in different industry sectors during our testing period. Most notably, for sic7 and sic9 industry, principal component regression model has CPSA over others for certain state variables. Testing solely for SPA or MCS would not have been able to show this result.

Except for sic0 industry for $h = 1$, the ridge model rejects the CSPA null for all industries and horizons, similar to previous forecast evaluation results. Therefore, when using a large collection of predictors, we argue that reducing the number of parameters to increase the quality of estimation is more important that shrinkage. We suspect a repeated forecast exercise using a smaller set of predictors may show changes in performance of ridge model. We leave this for future research.

Table 7: By industry: h=1, CSPA results

		ADS	ANFCI	NFCI	rec prob	vxo	gecon	mpu	epu	JLN f	JLN r	JLN m
	0forecast											
	Enet	⋆	⋆	⋆	⋆	⋆	⋆		⋆	⋆	⋆	⋆
	LASSO											
sic0	PCR	⋆	⋆	⋆	⋆	⋆	⋆		⋆	⋆	⋆	⋆
	PLS	⋆	⋆	⋆		⋆	⋆			⋆	⋆	⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆		⋆	⋆	⋆	⋆
	RF		⋆			⋆		⋆				
	0forecast									⋆		
	Enet	⋆	⋆	⋆		⋆		⋆	⋆	⋆		⋆
	LASSO	⋆	⋆	⋆		⋆		⋆	⋆	⋆		⋆
sic1	PCR	⋆	⋆	⋆		⋆		⋆	⋆	⋆	⋆	⋆
	PLS		⋆	⋆	⋆	⋆		⋆		⋆	⋆	⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF											
	0forecast	⋆		⋆	⋆	⋆	⋆			⋆	⋆	⋆
	Enet			⋆		⋆				⋆		⋆
	LASSO			⋆		⋆				⋆		⋆
sic2	PCR	⋆	⋆	⋆	⋆	⋆	⋆			⋆	⋆	⋆
	PLS			⋆		⋆				⋆		⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF	⋆	⋆	⋆	⋆	⋆	⋆			⋆	⋆	⋆
	0forecast					⋆				⋆		
	Enet					⋆				⋆		
	LASSO					⋆				⋆		
sic3	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS					⋆				⋆		
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF					⋆				⋆	⋆	
	0forecast	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Enet	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	LASSO	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
sic4	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF											
	0forecast	⋆	⋆	⋆		⋆	⋆		⋆	⋆	⋆	⋆
	Enet	⋆	⋆	⋆		⋆			⋆	⋆	⋆	⋆
	LASSO	⋆	⋆	⋆		⋆			⋆	⋆	⋆	⋆
sic5	PCR	⋆	⋆	⋆		⋆	⋆		⋆	⋆	⋆	⋆
	PLS	⋆	⋆	⋆		⋆	⋆		⋆	⋆	⋆	⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF											
	0forecast	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Enet	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	LASSO	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
sic6	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF											
	0forecast	⋆		⋆		⋆			⋆	⋆	⋆	⋆
	Enet									⋆		
	LASSO	⋆			⋆		⋆			⋆		
sic7	PCR	⋆		⋆		⋆	⋆			⋆	⋆	⋆
	PLS											
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	0forecast	⋆	⋆	⋆	⋆	⋆		⋆		⋆	⋆	⋆
	Enet	⋆	⋆	⋆		⋆				⋆		⋆
	LASSO	⋆	⋆	⋆		⋆				⋆		⋆
sic8	PCR	⋆	⋆	⋆		⋆		⋆	⋆	⋆		⋆
	PLS	⋆	⋆	⋆		⋆				⋆		⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF				⋆	⋆				⋆		⋆
	0forecast	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Enet											
	LASSO			⋆						⋆		
sic9	PCR										⋆	⋆
	PLS											
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆		⋆	⋆

A model that reject CSPA null with 5% significance marked (⋆).

Table 8: By industry: h=3, CSPA results

		ADS	ANFCI	NFCI	rec prob	vxo	gecon	mpu	epu	JLN f	JLN r	JLN m
	0forecast				⋆							⋆
	Enet	⋆	⋆	⋆		⋆				⋆	⋆	⋆
	LASSO	⋆	⋆	⋆		⋆				⋆	⋆	⋆
sic0	PCR	⋆	⋆	⋆		⋆		⋆	⋆	⋆	⋆	⋆
	PLS											
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF		⋆									
	0forecast									⋆		
	Enet									⋆		
	LASSO									⋆		
sic1	PCR	⋆	⋆	⋆		⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS	⋆			⋆		⋆			⋆		
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF	⋆	⋆	⋆	⋆		⋆					
	0forecast			⋆		⋆	⋆			⋆		⋆
	Enet			⋆		⋆	⋆			⋆		⋆
	LASSO			⋆		⋆	⋆			⋆		⋆
sic2	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆		⋆		⋆
	PLS			⋆		⋆	⋆			⋆		⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF	⋆	⋆	⋆	⋆					⋆		
	0forecast		⋆	⋆		⋆		⋆	⋆	⋆	⋆	⋆
	Enet					⋆				⋆		
	LASSO					⋆				⋆		⋆
sic3	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS					⋆				⋆		⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF	⋆			⋆					⋆	⋆	⋆
	0forecast	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Enet	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	LASSO	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
sic4	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF											
	0forecast	⋆	⋆	⋆		⋆	⋆			⋆		⋆
	Enet	⋆	⋆	⋆		⋆	⋆			⋆		⋆
	LASSO	⋆	⋆	⋆		⋆	⋆			⋆		⋆
sic5	PCR	⋆	⋆	⋆		⋆	⋆			⋆		⋆
	PLS	⋆	⋆	⋆		⋆	⋆			⋆		⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF											
	0forecast	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Enet	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	LASSO	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
sic6	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF											
	0forecast	⋆				⋆			⋆	⋆		
	Enet					⋆				⋆		
	LASSO					⋆			⋆	⋆		
sic7	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆		⋆	
	PLS											
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	0forecast	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Enet	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	LASSO		⋆			⋆				⋆	⋆	⋆
sic8	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS					⋆				⋆		⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF	⋆	⋆	⋆	⋆	⋆	⋆			⋆		
	0forecast	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Enet					⋆		⋆	⋆	⋆		
	LASSO					⋆				⋆		
sic9	PCR	⋆		⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS					⋆						
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF	⋆	⋆	⋆	⋆	⋆	⋆			⋆	⋆	⋆

A model that reject CSPA null with 5% significance marked (⋆).

Table 9: By industry: h=6, CSPA results

		ADS	ANFCI	NFCI	rec prob	vxo	gecon	mpu	epu
	0forecast	⋆							
	Enet		⋆	⋆					
	LASSO		⋆	⋆	⋆				
sic0	PCR	⋆	⋆	⋆		⋆			
	PLS		⋆	⋆	⋆				
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF								
	0forecast								
	Enet								
	LASSO					⋆			
sic1	PCR								
	PLS								
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF								
	0forecast			⋆		⋆			
	Enet			⋆		⋆	⋆		
	LASSO			⋆		⋆	⋆		
sic2	PCR			⋆		⋆	⋆		
	PLS			⋆		⋆			
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF								
	0forecast		⋆			⋆			⋆
	Enet					⋆			
	LASSO					⋆			
sic3	PCR		⋆	⋆	⋆	⋆		⋆	
	PLS					⋆			
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF								
	0forecast	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Enet	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	LASSO	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
sic4	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF								
	0forecast			⋆		⋆	⋆		
	Enet	⋆	⋆	⋆		⋆	⋆		
	LASSO	⋆	⋆	⋆		⋆	⋆		
sic5	PCR		⋆	⋆		⋆	⋆		
	PLS	⋆	⋆	⋆		⋆	⋆		
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF								
	0forecast	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Enet	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	LASSO	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
sic6	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF								
	0forecast					⋆		⋆	
	Enet					⋆			
	LASSO					⋆			
sic7	PCR					⋆		⋆	
	PLS								
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	0forecast	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Enet	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	LASSO	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
sic8	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS					⋆			
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF								
	0forecast			⋆	⋆	⋆			
	Enet	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	LASSO	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
sic9	PCR			⋆		⋆			
	PLS					⋆			
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF	⋆	⋆	⋆	⋆	⋆	⋆		

A model that reject CSPA null with 5% significance marked (⋆).

Table 10: By industry: h=12, CSPA results

		ADS	ANFCI	NFCI	rec prob	vxo	gecon	mpu	epu	JLN f	JLN r	JLN m
	0forecast	⋆			⋆		⋆		⋆			⋆
	Enet	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	LASSO	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
sic0	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS											
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF											
	0forecast											
	Enet		⋆	⋆	⋆						⋆	⋆
	LASSO				⋆							
sic1	PCR	⋆	⋆	⋆						⋆		⋆
	PLS											
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF						⋆					
	0forecast	⋆	⋆	⋆		⋆				⋆	⋆	⋆
	Enet		⋆	⋆		⋆				⋆	⋆	⋆
	LASSO		⋆	⋆		⋆				⋆	⋆	⋆
sic2	PCR		⋆	⋆		⋆	⋆			⋆	⋆	⋆
	PLS		⋆	⋆		⋆				⋆	⋆	⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF											
	0forecast		⋆	⋆		⋆	⋆	⋆		⋆		⋆
	Enet			⋆		⋆				⋆		
	LASSO			⋆		⋆				⋆		
sic3	PCR		⋆	⋆	⋆	⋆	⋆	⋆		⋆	⋆	⋆
	PLS			⋆		⋆				⋆		
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF									⋆		
	0forecast	⋆	⋆	⋆	⋆	⋆	⋆	⋆		⋆	⋆	⋆
	Enet	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	LASSO	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
sic4	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF											
	0forecast	⋆		⋆		⋆		⋆		⋆	⋆	⋆
	Enet	⋆	⋆	⋆		⋆	⋆	⋆				⋆
	LASSO	⋆	⋆	⋆		⋆	⋆	⋆				⋆
sic5	PCR	⋆	⋆	⋆		⋆				⋆		⋆
	PLS	⋆	⋆	⋆		⋆	⋆	⋆			⋆	⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF											
	0forecast	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Enet	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	LASSO	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
sic6	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF											
	0forecast					⋆						
	Enet					⋆				⋆		
	LASSO					⋆				⋆		
sic7	PCR					⋆			⋆			⋆
	PLS											
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	0forecast	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Enet		⋆	⋆		⋆			⋆			
	LASSO		⋆	⋆		⋆			⋆			
sic8	PCR	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	PLS					⋆						
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF											
	0forecast	⋆							⋆		⋆	
	Enet		⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	LASSO	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
sic9	PCR											
	PLS											
	Ridge	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	RF					⋆				⋆		

A model that reject CSPA null with 5% significance marked (⋆).

6 Conclusion

We investigate the predictive ability of supervised learning models predicting individual firm excess returns. Our main contribution is forecast evaluation of models for their relative predictive ability. We evaluate models forecasting all individual firm returns, as well as for each group of firms belonging to different industry sectors in the US. First, using tests of SPA, CSPA, and constructing MCS, we find that partial least squares and random forest perform better than others when forecasting for all firms in our data. Second, when models forecast returns for each industry sector, there is no single model that has SPA, CSPA over other models and is always included in MCS. We argue that this is due to the heterogeneity of firms across different industry sectors.

For future research, we aim to include more supervised machine learning models for forecast evaluation and utilize CSPA further to search for more state dependency in models' predictive performance, by using a different set of predictors or forecasting other time periods.

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.

Politis, D. N., and J. P. Romano (1994): "The stationary bootstrap," *Journal of the American Statistical association*, 89(428), 1303–1313.

Tibshirani, R. (2011): "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3), 273–282.

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.

Zou, H., and T. Hastie (2005): "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, 67(2), 301–320.