The Best of Both Worlds:

Forecasting US Equity Market Returns using a Hybrid Machine Learning – Time Series Approach*

Haifeng Wang, Harshdeep Singh Ahluwalia, Roger A. Aliaga-Díaz, and Joseph H. Davis†

This version: December 2, 2019

ABSTRACT

Predicting long-term equity market returns is of great importance for investors to strategically allocate their assets. We apply machine learning methods to forecast 10-year-ahead U.S. stock returns and compare the results to traditional Shiller regression-based forecasts more commonly used in the asset-management industry. Machine-learning forecasts have similar forecast errors to a traditional return forecast model based on lagged CAPE ratios. However, machine-learning forecasts have *higher* forecast errors than the regression-based, two-step approach of Davis et al [2018] that forecasts the CAPE ratio based on macroeconomic variables and then imputes stock returns. When we combine our two-step approach with machine learning to forecast CAPE ratios (a hybrid ML-VAR approach), U.S. stock return forecasts are statistically and economically more accurate than all other approaches. We discuss why and conclude with some best practices for both data scientists and economists in making real-world investment return forecasts.

Keywords: Machine learning, stock return forecasting, return predictability, CAPE ratio

* We wish to acknowledge Andrew Clarke, Douglas Grim, Dhagash Metha, and Edward Mitby for their useful comments. The views expressed in this working paper do not necessarily reflect the views of The Vanguard Group, Inc.

† All authors are with Vanguard Investment Strategy Group

JEL-Classification:	G10, C58, E37
First Draft:	December 2, 2019
This Version:	December 2, 2019

INTRODUCTION

Predicting long-term equity market returns is of great importance for investors to plan their strategic asset allocation. The cyclically adjusted price/earnings (P/E) ratio, or Shiller's CAPE ratio (Campbell and Shiller [1988, 1998]), is one of the most widely followed equity market valuation metrics in the investment profession. A high CAPE ratio has been associated with below average 10-year-ahead U.S. stock returns and vice-versa. Typically, researchers express this relationship in terms of a linear regression (Shiller's regression) of 10-year-ahead equity returns on the beginning period's CAPE ratio.

However, the accuracy of the Shiller regression has deteriorated since the 2000s. Davis et al [2018] suggest that a structural break around the 1990s dot-com bubble may have rendered that historical statistical relationship between CAPE and equity returns unstable. For instance, the average CAPE ratio for 2000-2019 is almost twice as high as the average pre-2000s and CAPE has not reverted back to its sample mean since the 1990s. Recent studies such as Siegel [2016] and Philips and Ural [2016] have attempted to reformulate the construction of the Shiller CAPE ratio around this structural break with mixed results.

Arnott et al [2017] show CAPE, interest rate, and inflation are interlinked. Moderate levels of inflation and real interest rates seem to allow high valuation multiples. Valuation does not always predict stock market returns, because the "normal" level of CAPE is non-stationary and will vary with changing economic conditions. Similarly, Davis et al [2018] attribute the lack of mean reversion in CAPE to the low real interest rates and inflation of the last two decades. Thus, Davis et al [2018] propose a two-step approach that conditions mean reversion in the CAPE ratio on real bond yields, inflation, and financial volatility. This approach shows higher forecast accuracy than the original Shiller regression.

The success (and limitations) of traditional statistical approaches, whether Shiller's original formulation or any other approach, depend on an analyst's ability to identify the right variables and the right statistical specification for the forecasting regression. In general terms, machine learning (ML) approaches, in general terms, can assess a large number of combinations of forecasting variables, including non-linear terms. More importantly, Many ML algorithms do not require a pre-specified regression equation (or functional form) in order to forecast equity returns. This flexibility means that ML algorithms can, in theory, illuminate true but unknown data generating processes for stock returns that would be missed in a linear forecasting regression.

In this paper, we analyze whether different ML techniques can indeed improve on the accuracy of long-run U.S. stock market returns forecasts out-of-sample. To our knowledge, this is the first study to systematically assess if ML techniques can outperform the traditional Shiller's CAPE regression and other statistical methods such as Davis et al [2018]. We find that some ML models produce marginally lower forecast errors than the traditional Shiller regression, but none of them are statistically better than a naïve forecast based on historical average stock returns. None of the ML algorithms can improve on the forecast accuracy of the two-step approach developed by Davis et al [2018].

We then turn to the question of whether a hybrid approach that combines ML algorithms and the two-step statistical formulation (Davis et al [2018]) can attain results better than some other alternative proposed in academic literature. More specifically, instead of using ML techniques to forecast equity returns directly, we specify a hybrid ML-VAR to forecast the earnings yield (1/CAPE) along with other macroeconomic variables, similar to a traditional vector autoregressive model (VAR). In a second step, following Davis et al [2018], we calculate future stock returns from the forecasted CAPE ratios. We find the forecast errors of this hybrid approach to be much lower than those of the Shiller regression or any ML algorithm in isolation. Moreover, the hybrid approach, because of its ability to capture non-linear relationships, also produces forecasts that are superior to the original Davis et al [2018] VAR-based two-step approach.

We organize the paper as follows. We first briefly review the fast-growing literature of ML applications in empirical asset pricing and financial forecasting. Second, we discuss the challenges in applying ML models to long-term stock prediction and offer ways to address the challenges. In the third section, we compare Shiller's forecasting regression with four of the most prominent ML techniques. In the fourth section, we explore the performance of a hybrid approach, leveraging the power of ML techniques, but within the more structured framework developed in Davis et al [2018]. In the final section we offer some recommendations based on our findings. We find applying ML techniques within a robust economic framework to be superior to applying such techniques in isolation. Practitioners in finance can benefit tremendously from exploring the intersecting realms of empirical asset pricing and machine learning.

Section 1 – Machine Learning and Asset Pricing

The dramatic rise in computing power, sophisticated yet easily usable ML algorithms, and the availability of large quantities of data have made ML applications much more accessible to financial forecasters. ML is a broad, catchall concept that leverages algorithms and statistical models, thereby enabling computers to learn patterns in historical data and forecast based on the learning. As ML techniques have advanced, researchers have begun to compare their predictive power with that of linear regression, a widely used traditional statistical technique.

ML techniques differ from linear regression in many ways. Most importantly, unlike linear regression where parametric formulation is known and fixed, ML techniques are much more flexible. Many of these ML techniques rely on algorithms to recognize patterns, test the validity of these patterns, and make predictions based on patterns that are considered systematic (as opposed to "patterns" that are just random), regardless of any functional constraints. However, both linear regression and traditional ML techniques share a similar structure, with independent or explanatory variables on the right-hand side of the equation and the dependent or predicted variable on the left-hand side. Because of the explicit input-output pairs, ML techniques are also considered to be supervised machine learning models.

A more computationally intensive form of ML technique is deep learning (DL). Under DL, the layers of features are not designed by humans. Instead, they are learned from data using a general-purpose learning procedure. DL discovers intricate structures in data by using algorithms to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer.

In this paper, we use the term "ML" to describe both traditional ML and DL techniques. We consider the application of Random Forest (RF), Gradient Boosting Machine (GBM), Support Vector Regression (SVR), and Gated Recurrent Units (GRU) in predicting long-term stock market returns.

RF is a modified and improved version of bootstrap aggregation (bagging), which aims to reduce the variance of an estimated prediction function through averaging all predictions. A RF model consists of many random decision trees. Two types of randomnesses are built into the model. First, each tree is built on a random sample from the original data. Second, at each tree node, a subset of features are randomly selected to generate the best split. These randomnesses allow RF models to build a large collection of de-correlated trees and average them, thus achieving a better results than bagging. Khaidem et al [2016] combine RF with momentum indicators such as Relative Strength Index and stochastic oscillator, and find substantial improvement in predicting short-term stock returns.

Unlike RF, where multiple, uncorrelated trees are built, GBM builds an ensemble of shallow and weak successive trees with each tree learning and improving on the previous. When combined, these many weak successive trees produce a powerful “wisdom of the algorithmic cloud” that often significantly improve the prediction. In other words, GBM first models data with simple models and analyzes data for errors. These errors signify data points that are difficult to fit by a simple model. GBM then sequentially focuses on those hard to fit data and addresses them. The model combines all the predictors by giving some weights to each predictor. Using GBM to improve forecast accuracy while forecasting one month ahead U.S. stock returns, Krauss et al (2017) found GBM contributes to positive excess returns when implementing an arbitrage strategy on S&P500 index.

Different from tree-based models such as RF and GBM, the objective of an SVR algorithm is to find a hyperplane in an N-dimensional space (N is the number of features) that distinctly separates the data points. There are many possible hyperplanes that could be chosen in an N-dimensional space. SVR attempts to find a plane that maximizes a margin, such as distance, between data points in both classes, providing some reinforcement so that future data points can be classified with more confidence. Kim [2003] finds SVR to be a useful tool in predicting the sign of daily returns for the Korea composite stock price index.

GRU aims to solve the vanishing gradient problem that arises in a standard recurrent neural network. The vanishing gradient problem means that as a model goes back to the lower layers, weights may never change at these lower layers. This leads to the output being mostly affected by the value close to it. To address this, GRU uses two vectors, an update gate and a reset gate, to decide what information should be passed to the output. These special features allow the model to be trained to keep information from long ago, without washing it through time, or to remove information which is irrelevant to the prediction. Minh et al [2018] predict the direction of S&P500 index by applying financial news and sentiment to GRU. They find GRU correctly predicted the direction of the index two thirds of the time.

It is worth noting that despite its promise, ML is not without flaws, especially in the real application to money management (López de Prado [2018]). The features of ML algorithms, the flexibility of parameters available to tune in ML models (or so-called hyperparameters), and the growth of computing power become a double-edged sword. If designed and implemented poorly, ML algorithms suffer from look-ahead bias and data mining (López de Prado [2019], Zhang [2007]). Another drawback of some ML models is the difficulty of interpretability (Ribeiro et al, [2016]; Doshi-Velez and Kim, [2017]). Although work is well underway to link statistical inference with ML (see, for example, Wager and Athey, [2018]) and causal effects (Athey and Imbens, [2017]), some techniques, especially those associated with neural networks, still appear to be opaque, hindering their acceptance and adoption.

Section 2 – Addressing data and look-ahead bias issues in ML

To date, most literature in applying ML to market prediction focuses on short-term returns and often involves a cross-section of stocks, partly to increase the sample size as ML is very data-intensive to train and partly to take advantage the resampling and cross-validation techniques established in other ML applications (i.e. drug testing or credit card default). However, it is challenging to apply ML to predict long-term returns because of data availability.

In this paper, we offer three solutions to address this issue by combing data science with domain knowledge, while addressing look-ahead bias that has become a focal point in industry debate. First of all, we adjust hyperparameters in-sample. For example, we examine a range of learning rates in GBM. A small learning rate enables GBM to train the model more slowly with more data, while a larger learning rate trains the model quickly but may be less accurate. We vary epoch and batch size in GRU to test hyperparameters fit for our sample size. We attempt to address any accompanying overfitting issues by keeping prediction strictly out-of-sample. For example, we do not try different iterations until the model looks good both in-sample and out-of-sample, as alerted in Arnott et al [2019]. Rather, we train the model in-sample with different hyperparameter combinations. We examine the training results, pick the hyperparameter combination with the lowest RMSE, and apply the combination to the out-of-sample data. In other words, we use the out-of-sample data only once for each ML model. Exhibit 1 shows the hyperparameters tuned in this paper.[1]

Secondly, we use a recursive window. As we train the models, we recursively estimate models by expanding the training data window one month at a time. After training the model, we choose the hyperparameters with the lowest RMSE and apply these hyperparamters out-of-sample. We still use the recursive window for out-of-sample forecasts, meaning each prediction will have the same hyperparameters but different parametric formulation. The parametric formulation to predict the next month is only determined by data up to the last month. In other words, we retrain the model every time there is a new observation until the second to last observation, giving the maximum flexibility to every ML technique and expanding available data sample.[2]

[1] We detail how we divide training and testing data in Appendix

[2] The exception is the GRU approach. It would be extremely time-consuming to build and train these many GRU.

Exhibit 1 *Hyperparameters assessed in this analysis*

Gradient Boosting Machine	**Random Forest**	**Support Vector Machine**	**Gated Recurrent Units**[3]
n_estimator	n_estimator	C	learning_rate
max_depth	max_depth	epsilon	epoch
learning_rate	min_samples_split	kernel	batch_size
125 combinations	125 combinations	100 combinations	80 combinations

Last but not least, we limit the independent variables to those we believe are more effective to explain long-term stock returns based on literature. Econometric models often require a certain number of observations to draw statistically meaningful inference. For example, the 10-1 rule of thumb means for every variable, there should be at least ten observations (van Smeden et al [2018]). Such rules hold less ground in ML as the number of trees to grow or the number of branches to have for each tree will have different implication for the sample size. In GRU, the node and layer of the neural net will require exponentially higher numbers of observations relative to independent variables (Du et al, 2018). Therefore, we are selective in choosing independent variables, as explained in the following sections.

Section 3- Applying Machine Learning to Shiller's forecasting regression

Over time, long-term U.S. stock returns have tended to move inversely to beginning-of-period CAPE ratios. Based on this observation, financial analysts express 10-year-ahead stock returns as a linear function of the latest Shiller CAPE ratio via a simple linear regression, commonly called the Shiller Regression:

X Y

$\boxed{C_t}$ — $\boxed{R_{t+120}}$

[3] We use 1 layer of neural net in GRU

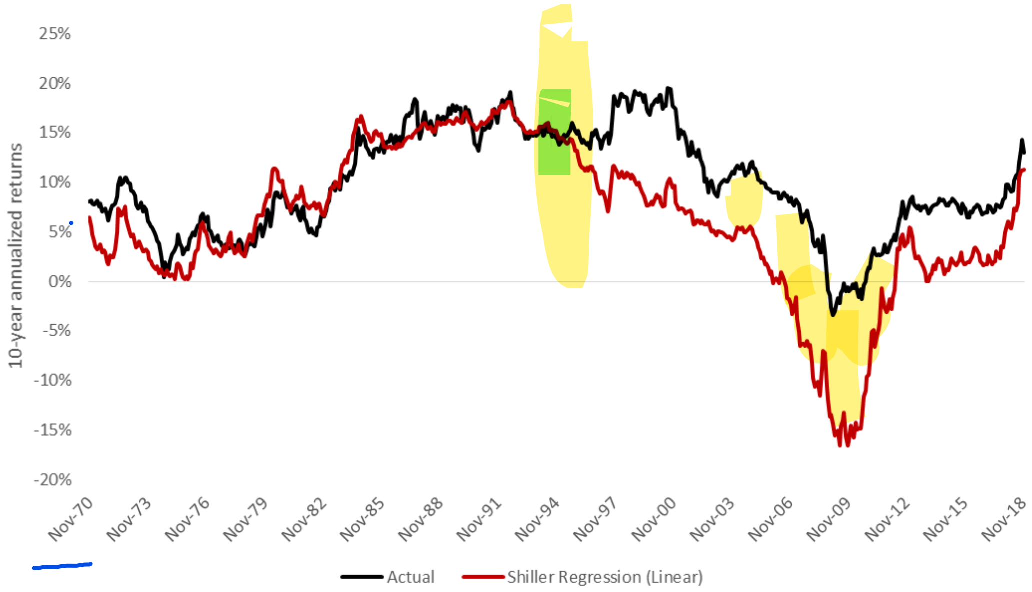
195

(1) $$\left[\underline{R_{t+120}} = \alpha + \beta \, \underline{CAPE_t} + \epsilon_t . \right]$$

The CAPE ratio has explained a remarkably high 54% of the time-series variation in 10-year-ahead nominal U.S. stock returns, as measured by Equation (1)'s in-sample, or fitted, R^2, over the 1926-2018 period.

Unfortunately, the Shiller regression's out-of-sample forecast accuracy has weakened since the mid-1980s. Exhibit 2 illustrates that beginning with the forecasts made in this period, regression-projected stock returns have generally been too bearish, even when one includes the 1999 tech bubble. Put another way, over the last few decades, real-time investors would have been better served by using the historical average returns (Davis et al [2018]).

Exhibit 2 *The CAPE Ratio's Predictive Power has diminished*



Notes: For the real-time analysis, the regression coefficients are determined recursively at a monthly frequency, starting with January 1926 – December 1959 data and re-estimating the regression coefficients every month thereafter. The gap between the two lines represents forecast error.

Source: Author's calculations. For details, refer to - Improving U.S. stock return forecasts: A "fair-value" CAPE approach (Davis, Aliaga-Diaz, Ahluwalia and Tolani, 2018), Journal of Portfolio Management, Vol. 44, No. 3 (2018), pp. 43-55. © 2018 Institutional Investors LLC. All rights reserved. http://jpm.iijournals.com/content/44/3/43

The degradation of its predictive power is partly due to the lack of mean reversion in the CAPE ratio itself. Both Siegel and Shiller attribute the elevated CAPE to an economic environment of low interest rates (Siegel, [2017], Shiller, [2017]). If changes in long-term real interest rates influence the fair-value of the CAPE ratio to which stock returns should revert, then the coefficients in traditional CAPE regressions will suffer from instability whenever there are meaningful changes in the level of real bond yields. Arnott et al [2017] show strong evidence that the CAPE mean reversion level varies over time with macroeconomic conditions. Davis et al [2018] propose that CAPE is dependent on the state of the economy as measured by real interest rates, inflation, and measures of financial volatility. Other literature, such as Welch and Goyal

(2007), examines the impact of a slew of variables, including stock variance, Treasury bill rates, long-term bond yields, and the default yield spread, on the equity premium. They found a few variables in some specifications that predict the long-term equity premium but most other models are not performing well.

Motivated by this literature, we extend the Shiller univariate regression to a multivariate regression, including CAPE, real interest rates, inflation, measures of financial volatility, stock variance, Treasury bill rates, the default yield spread, and the default return spread. We use 10-year ahead return as the predicted variable (Equation 2). Put simply, we attempt to forecast returns directly, just like the Shiller regression, but add a few important economic and market variables to the regression.

$$R_{t+120} = f(CAPE_t, Y_t, CPI_t, SPVOL_t, BondVol_t, SVAR_t, TBL_t, DFY_t, DFR_t)$$

Where

- CAPE is the cyclically adjusted price/earnings (P/E) ratio
- Y is Real 10-year bond yields, or nominal Treasury yield less an estimated 10-year expected inflation rate
- CPI is Year-over-year CPI inflation rate
- SPVol is the Realized S&P500 price volatility, over trailing 12 months
- BondVol is the Realized volatility of changes in our real bond yield series, over trailing 12 months
- SVAR is the stock variance computed as sum of squared daily returns on S&P 500
- TBL is the treasure bill rates
- DFY is the default yield spread computed as the difference between BAA- and AAA-rated corporate bond yields
- DFR is the default return spread, computed as the difference between the return on long-term corporate bonds and returns on the long-term government bonds

Exhibit 3 shows that the out-of-sample RMSE of the multiple regression comes out to be 6.6% while that of a naïve historical average forecast is 5.7%. Thus adding other important variables to a regression alone does not improve the forecasts.

Exhibit 3: *Comparison of real time predictive power of nominal U.S. stock market returns using Root Mean Square Error (RMSE)*

	MULTIPLE REGRESSION	**TWO STEP APPROACH**
HISTORICAL AVERAGE	5.7%	
LINEAR/VAR	6.6%	3.8%***
GBM	5.7%	
RF	6.2%	
SVR	6.4%	
GRU	7.2%	
ENSEMBLE 1	4.7%*	
ENSEMBLE 2	4.7%*	

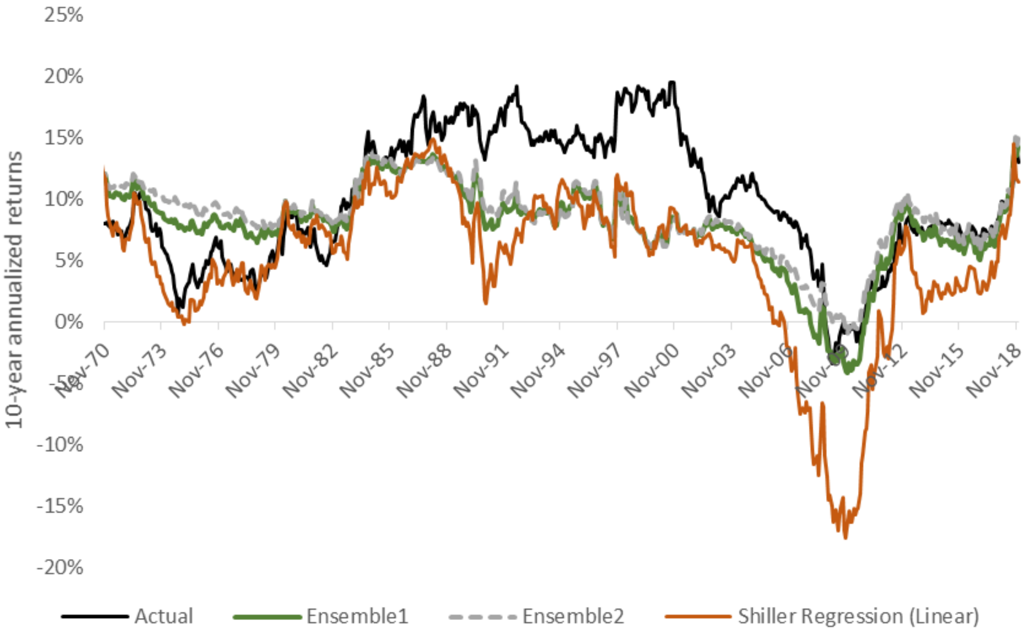
Notes: For the real-time analysis, the regression coefficients are determined recursively at a monthly frequency, starting with January 1926 – December 1959. A "" next to the RMSE refers to the significance (Newey-West adjusted) of the Diebold-Mariano test (2002) of whether the forecast is statistically better or worse than the historical mean. Significance level at 90%, 95% and 99% are denoted by one, two and three asterisks respectively. Source: Author's calculations*

The linearity assumption embedded in the multivariate regression is not the only cause of poor forecasts. We directly forecast returns using ML algorithms, based on the same predictors. Exhibit 3 shows that none of the individual models are statistically better than the naïve historical average forecast. Put another way, real-time investors would have been better served using the historical average return as the baseline for future stock returns. Strikingly, only some ML methods demonstrated a small edge in predictive power over the linear regression. GRU performed particularly poorly in terms of RMSE.

The inferior prediction results may be due to the following three reasons. Firstly and probably the most importantly, the ML methods outperform in-sample, but the hyperparameters trained in-sample do not fit data out-of-sample. Secondly, while we think the hyperparameters tested in GBM, RF and SVR encompass the most represented ones, the GRU have so many hyperparameters to tune that our analysis may not cover all of them. Last but not least, the limited amount of training data is worth noting. In particular, in calculating 10-year ahead returns, 10 years of data (or 120 few monthly data points) is sacrificed. Predicting 10-year ahead returns directly exacerbates the issue of data availability.

Despite the poor predictive power of individual models, we find an equally weighted combination of the all the ML model forecasts (Ensemble 1) and a combination of all the ML models and the traditional multiple regression (Ensemble 2) show a modest statistical improvement over the naïve forecast. Revealingly, both ensemble methods have lower RMSE than any of the individual methods used to build these ensemble methods. This result suggests that no one prediction method dominates the others and that each method contributes to the simple average during a period of time, which is then picked up by ensemble methods (Exhibit 3 and Exhibit 4). This finding is in line with Krauss et al [2017], Ahmed et al [2007] and Makridakis et al [2018] who conclude that ensemble methods are typically among the best performing models.

Exhibit 4 *The predictive power of ensemble approaches*



Notes: For the real-time analysis, the regression coefficients are determined recursively at a monthly frequency, starting with January 1926 – December 1959. For each ML methods, we choose the hyperparameters with the lowest RMSE and re-estimate the regression coefficients every month thereafter using the chosen hyperparameters.

Source: Authors' calculation

Section 4- Combining ML with two-step approach: ML-VAR

P/E

E/P

To address some of the potential issues regarding Shiller Regression discussed in the previous sections, Davis et al [2018] propose a two-step framework to forecast long-run equity market returns. The two-step approach is based on a VAR model to forecast the inverse of the CAPE ratio itself as the first step and to impute returns from the CAPE ratio in the second.

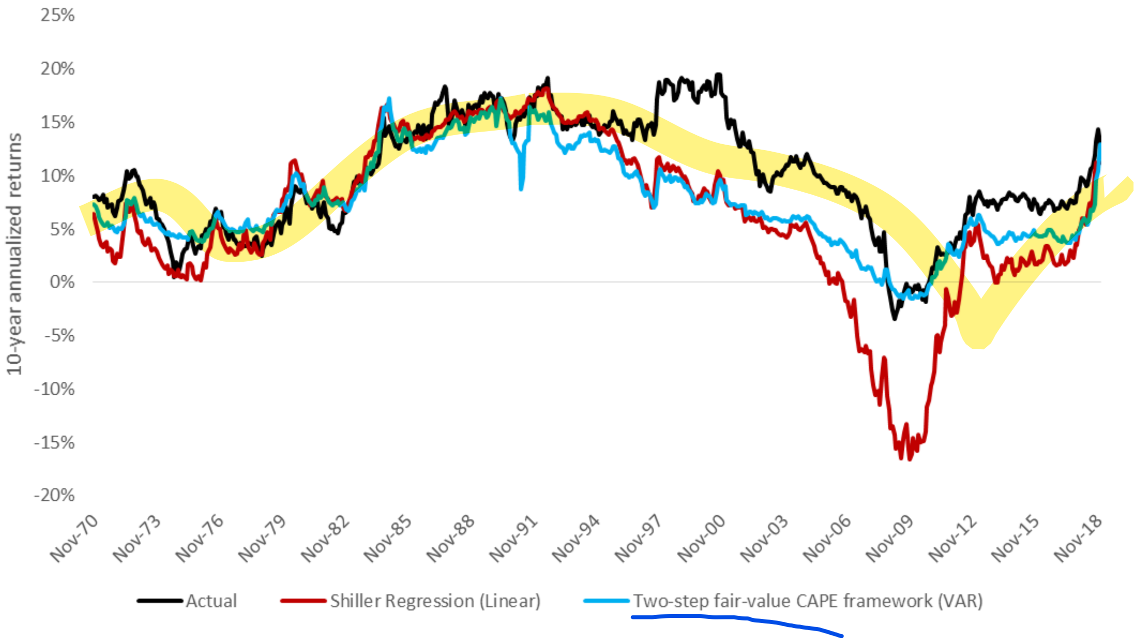
More specifically, step one estimates a VAR model with 12 monthly lags of 1/CAPE, real 10-year bond yields, CPI inflation rate, realized S&P500 price volatility, and realized volatility of changes in real bond yield. In the second step, stock returns are imputed as a sum of three parts: valuation expansion; earnings growth; and dividend yield, adapted from the framework of Bogle and Nolan [2015] and Ferreira and Santa-Clara [2011].[4] At any one point in time, the VAR forecasts the CAPE earnings yields out for 10 years, and step two derives the expected future 10-year-ahead return on U.S. stocks.

One benefit of the two-step approach is that the fair-value CAPE ratio – to which the actual CAPE ratio should revert – is permitted to vary over time, conditional on the movements of the other fundamental variables. Another benefit of using a VAR model, is that more data is available compared to forecasting 10-year ahead returns directly, an important advantage for ML methods in the following discussion.

Exhibit 5 shows the actual real-time forecast of the two-step model, which tracks actual returns fairly well, declining throughout the 2000s and anticipating a strong rebound immediately following the global financial crisis in 2009. With a RMSE of 3.8%, the traditional two-step model performs statistically better than the ML methods discussed previously.

[4] The benefit of the sum of parts approach is that it should mitigate so-called Stambaugh [1999] bias that can plague predictive regressions with persistent regressors like CAPE ratios that involve overlapping data (Nelson and Kim [1993])

Exhibit 5: *Two-step Fair-Value CAPE Model – Reasonable Out-Of-Sample Performance*



Notes: For the real-time analysis, the regression coefficients are determined recursively at a monthly frequency, starting with January 1926 – December 1959 data and re-estimate the regression coefficients every month thereafter. The gap between the two lines represents forecast error.

Source: Author's calculations. For details, refer to - Improving U.S. stock return forecasts: A "fair-value" CAPE approach (Davis, Aliaga-Diaz, Ahluwalia and Tolani, 2018), Journal of Portfolio Management, Vol. 44, No. 3 (2018), pp. 43-55. © 2018 Institutional Investors LLC. All rights reserved. http://jpm.iijournals.com/content/44/3/43

Building on the two-step framework above, we propose a refined framework that integrates ML methods with VAR to forecast the inverse of CAPE ratio. Importantly, we replace the linear core within the VAR with ML and forecast 1/CAPE, inflation, real yields, equity volatility, and bond volatility dynamically in a vector (see Equation 3). We then use the same sum-of-parts identity used in Davis et al [2018] (Equation 4).

$$(3)\ X_t = f(X_{t-1}, X_{t-2}, \dots, X_{t-12})$$

where $\boldsymbol{X_t}$ is a vector of the five variables in the logarithmic form, including:

- CAPE real earnings yield, or 1/CAPE

- Real 10-year bond yields, or nominal Treasury yield less an estimated 10-year expected inflation rate
- Year-over-year CPI inflation rate
- Realized S&P500 price volatility, over trailing 12 months
- Realized volatility of changes in our real bond yield series, over trailing 12 months

$$(4) \quad r_{t+1} = \%\Delta PE_{t+1} + \%\Delta E_{t+1} + DP_{t+1}$$

where %ΔPE is the percentage change in P/E ratio (or valuation expansion), %ΔE is earnings growth, and DP is the dividend yield.

The VAR model's forecast for the earnings yield provides the percentage change in CAPE ratios, %ΔPE, which is then added to the historical average of earnings growth and dividend yield. Similar to the multivariate regression, we also created two ensemble methods. We use the same training and testing data, hyperparameter combinations, and parameter tuning processes as used in section 3.

The result is shown in Exhibit 6. The right column shows that all ML-VAR methods demonstrate remarkable improvement and have statistically lower average errors than the naïve historical average forecast. The ML-VAR methods also have lower forecast errors than the original two-step approach. We see the highest improvement (RMSE drops the most) in GRU where the RMSE improves to 2.6%. The RMSE of Ensemble 1 (2.6%) is marginally lower than GRU and that of Ensemble 2 (2.8%) closely trail GRU. More important, applying the robust two-step framework with ML algorithms drastically improves the forecast accuracy of all non-linear ML techniques relative to the forecasts from those same ML techniques used without the two step framework

Exhibit 7a and 7b show the largest forecast error between predicted and actual returns occurred during the tech bubble. On the other hand, GRU differs by fitting well during the tech bubble and over-predicted returns after 2005. In other words, the lower RMSE of GRU comes almost entirely from accuracy during the tech bubble, arguably a period of irrational exuberance that a rational model should not be expected to predict well. This could be due to the methodological difference in our forecasting method for GRU, where this model was only estimated once in the training data set. Other ML models were recursively estimated by expanding the training data window one month at a time.

Exhibit 6: *Comparison of real time predictive power, Nominal U.S. stock market returns*

	SHILLER REGRESSION	ML-VAR TWO STEP APPROACH
HISTORICAL AVERAGE	5.7%	
LINEAR/VAR	6.6%	3.8%***
GBM	5.7%	3.6%***
RF	6.2%	3.6%***
SVR	6.4%	3.6%***
GRU	7.2%	2.6%***
ENSEMBLE 1	4.7%*	2.6%***
ENSEMBLE 2	4.7%*	2.8%***

Notes: For the real-time analysis, the regression coefficients are determined recursively at a monthly frequency, starting with January 1926 – December 1959. A "" next to the RMSE refers to the significance (Newey-West adjusted) of the Diebold-Mariano test (2002) of whether the forecast is statistically better or worse than the historical mean. Significance level at 90%, 95% and 99% are denoted by one, two and three asterisks respectively. Source: Author's calculations*

Exhibit 7a *Out of sample prediction from the ML-VAR approach*

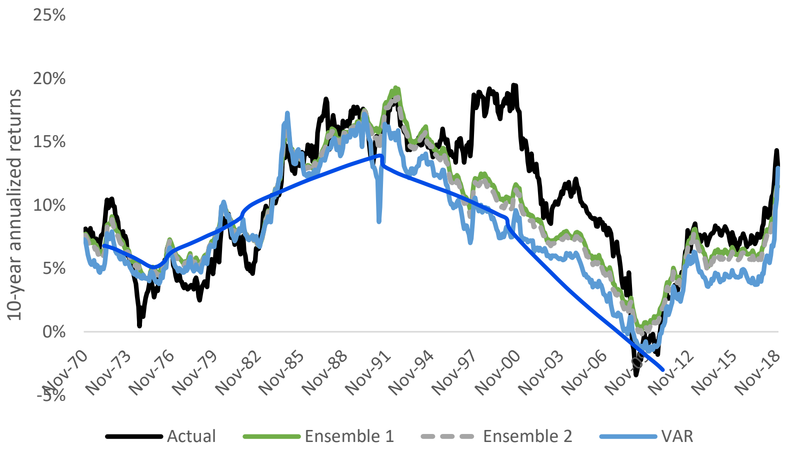
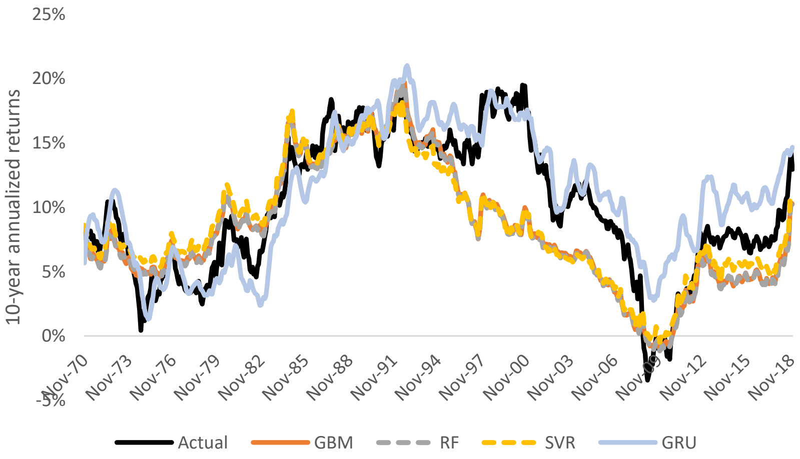


Exhibit 7b *Out of sample prediction from the ML-VAR approach*



Notes: For the real-time analysis, the prediction (with the exception of GRU) are determined recursively at a monthly frequency, starting with January 1926 – December 1959 data. For each ML methods, we choose the hyperparameters with the lowest RMSE and re-estimate the regression coefficients every month thereafter using the chosen hyperparameters. For GRU, the prediction are determined by running the entire training data. The chosen hyperparameters with the lowest RMSE are then applied to the GRU for the entire period.

Source: Authors calculation

Section 5- Conclusion

Valuation metrics such as P/E ratios are widely followed by the investment community because they are believed to predict future long-term stock returns. However, the out-of-sample accuracy of relying on the well-known Shiller CAPE regression to forecast stock returns is subpar. We show that adding important predictive variables and using ML methods to forecast returns directly in one step does not statistically improve the model's predictive power.

The Davis et al [2018] two-step approach, which forecasts CAPE conditional on the macroeconomic landscape, improves forecast accuracy significantly. In this article, we explore a novel approach to enhance the accuracy of the two-step model by combining it with non-linear ML techniques. Specifically, we forecast earnings yield using a hybrid ML-VAR approach to condition its mean reversion on the economy. Rather than simply using non-linear ML models to forecast only 1/CAPE, these algorithms are used to dynamically forecast five predictive variables (and their lagged values), similar to a VAR. We find that this ML-VAR approach lowers forecast errors over the full period from 1960 onwards by 50% for long horizon U.S. stock market returns.

The ensemble method, which averages all other model forecasts, consistently provides improved predictive power. This finding echoes conclusions made from a recent data science competition (with Ahmed et al [2007] and Makridakis et al [2018]). To date, however, the investment community has largely overlooked the potential benefits of ensemble models.

Overall, we find applying machine learning techniques within a robust economic framework (i.e. Davis et al [2018] two step approach) is far superior than applying such techniques in isolation (directly forecasting returns). We believe that both economists and data scientists will benefit tremendously from the intersecting realms of economics and machine learning.

References

Ahmed, Nesreen, Amir Atiya, Neamat Gayar and Hisham El-Shishiny "An Empirical Comparison of Machine Learning Models for Time Series Forecasting." *Econometric Reviews* 29(5-6) (2010) pp. 594-621

Arnott, Robert, Harvey Campbell, and Harry Markowitz "A Backtesting Protocol in the Era of Machine Learning." *The Journal of Financial Data Science*, 1 (1) (2019), pp. 64-74

Arnott, Robert, Denis B. Chaves and Tzee-man Chow "King of the Mountain: The Shiller P/E and Macroeconomic Conditions." *The Journal of Portfolio Management*, 44 (1) (2017), pp. 55-68

Athey, Susan and Guido Imens (2017) "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives*, Vol. 31(2), (2017), pp. 3-32

Bogle, J.C. "Investing in the 1990s: Occam's Razor Revisited." *Journal of Portfolio Management*, Vol. 18(1) (1991), pp. 88–91.

Bogle, John C. and Michael W. Nolan, Jr. "Occam's Razor Redux: Establishing Reasonable Expectations for Financial Market Returns." *Journal of Portfolio Management*, vol. 42(1), (2015), pp. 119-134

Campbell, John Y. and Robert J. Shiller "Stock Prices, Earnings, and Expected Dividends." *Journal of Finance*, vol. 43(3), (1988), pp. 661–676.

Campbell, John Y. and Robert J. Shiller "Valuation Ratios and the Long-Run Stock Market Outlook." *Journal of Portfolio Management*, vol. 24, no. 2 (1998) 11–26.

Davis, Joseph, Roger Aliaga-Díaz, Harshdeep Ahluwalia and Ravi Tolan "Improving U.S. Stock Return Forecasts: A "Fair-Value" CAPE Approach." *The Journal of Portfolio Management* (2017)

Diebold, Francis X., and Robert S. Mariano "Comparing Predictive Accuracy." *Journal of Business and Economic Statistics,* vol. 20, no. 1, (2002), pp. 134-144.

Doshi-Velez , Finale and Been Kim "Towards A Rigorous Science of Interpretable Machine Learning", arXiv:1702.08608, (2017)

Du, Simon, Yang Wang, Xiyu Zhai, Sivaraman Balakrishnan, Ruslan Salakhutdinov, and Aarti Singh "How Many Samples are Needed to Estimate a Convolutional or Recurrent Neural Network", arXiv:1805.07883, (2018)

Ferreira, Miguel, A. and Pedro Santa-Clara. "Forecasting Stock Market Returns: The Sum of the Parts is More than the Whole." *Journal of Financial Economics*, vol. 100, no. 3, (2011), pp. 514–537.

Goyal, A., and I. Welch. "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction." *The Review of Financial Studies*, Vol. 21, No. 4 (2008), pp. 1455-1508.

Gu, Shihao, Bryan Kelly, Dacheng Xiu "Empirical Asset Pricing via Machine Learning". (2018), available at: https://dachxiu.chicagobooth.edu/download/ML.pdf

Kim, Kyoung-jae “Financial time series forecasting using support vector machines.” *Neurocomputing*. (2003), pp: 307-319

Krauss, Christopher, Xuan Anh Do, and Nicolas Huck “Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500.” *European Journal of Operational Research* (2017) pp. 689-702

López de Prado, Marcos “The 10 reasons most machine learning funds fail”. *The Journal of Portfolio Management* (2018), pp. 120-133

López de Prado, Marcos “A Data Science Solution to the Multiple-Testing Crisis in Financial Research” *The Journal of Financial Data Science* (2019), pp. 99-110

Makridakis S, Spiliotis E, Assimakopoulos V “Statistical and Machine Learning forecasting methods: Concerns and ways forward.” PLoS ONE 13(3): e0194889. https://doi.org/10.1371/journal.pone.0194889

Ribeiro, Marco, Sameer Singh and Carlos Guestrin “Why Should I Trust You?” Explaining the Predictions of Any Classifier” (2016), available at: https://arxiv.org/abs/1602.04938

van Smeden, Maarten, Karel GM Moons, Joris AH De Groot, and Gary S Collins “Sample size for binary logistic prediction models: Beyond events per variable criteria.” *Statistical Methods in Medical Research* 28(8):096228021878472, (2018)

Wager, Stefan and Susan Athey “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests”, *Journal of the American Statistical Association*, 113:523, (2018), pp. 1228-1242, DOI: 10.1080/01621459.2017.1319839

Zhang, Peter “Avoiding Pitfalls in Neural Network Research” *IEEE Transactions on systems, ma, and cybernetic – Part C: Applications and Reviews*, Vol. 37, No.1, (2007)

Appendix

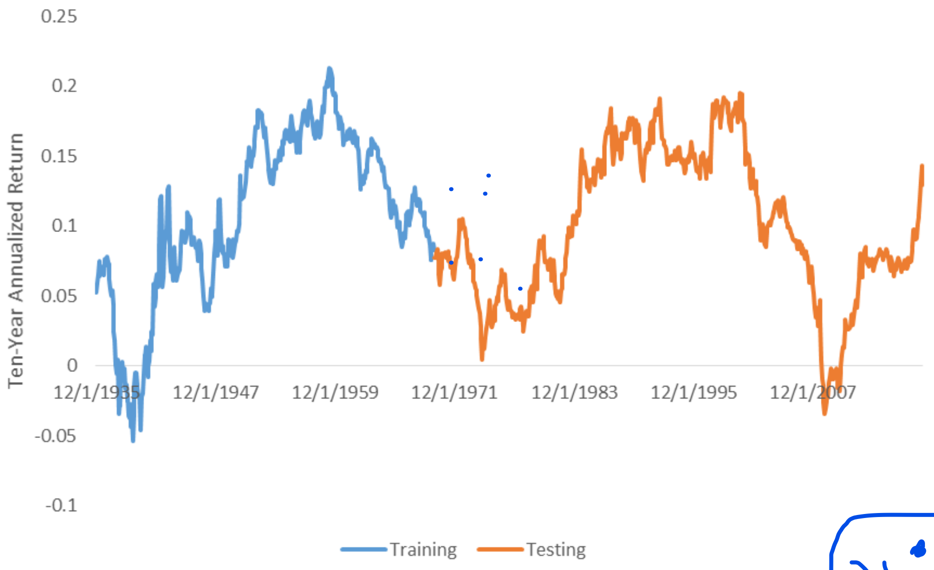
Data and out-of-sample forecasting

All the data in this article were obtained from Professor Shiller's website at aida.wss.yale.edu/~shiller/data.htm. Real bond yields reflect the nominal 10-year U.S. Treasury yield, less an estimate of 10-year ahead inflation expectations. A consistently defined and long-term running series on U.S. inflation since the 1920's does not exist.

Our synthetic inflation expectations series was derived so that an investor could have replicated them at the time our stock forecasts were made. Specifically, we defined inflation expectations as the average of the predicted CPI inflation rate over the next 10 years generated from an AR model at any month in time. The AR model included 12 monthly lags in annualized CPI inflation rates and was estimated using a 30-year rolling window. The synthetic time series for our expected 10-year inflation rate (Davis et al [2018]).

We divide our monthly sample data into training data and testing data to avoid look-ahead bias (Exhibit A1). The training data is from January 1926 – December 1959. The testing data is from January 1960 – December 2008.

Exhibit A1 *The Training and testing data split*



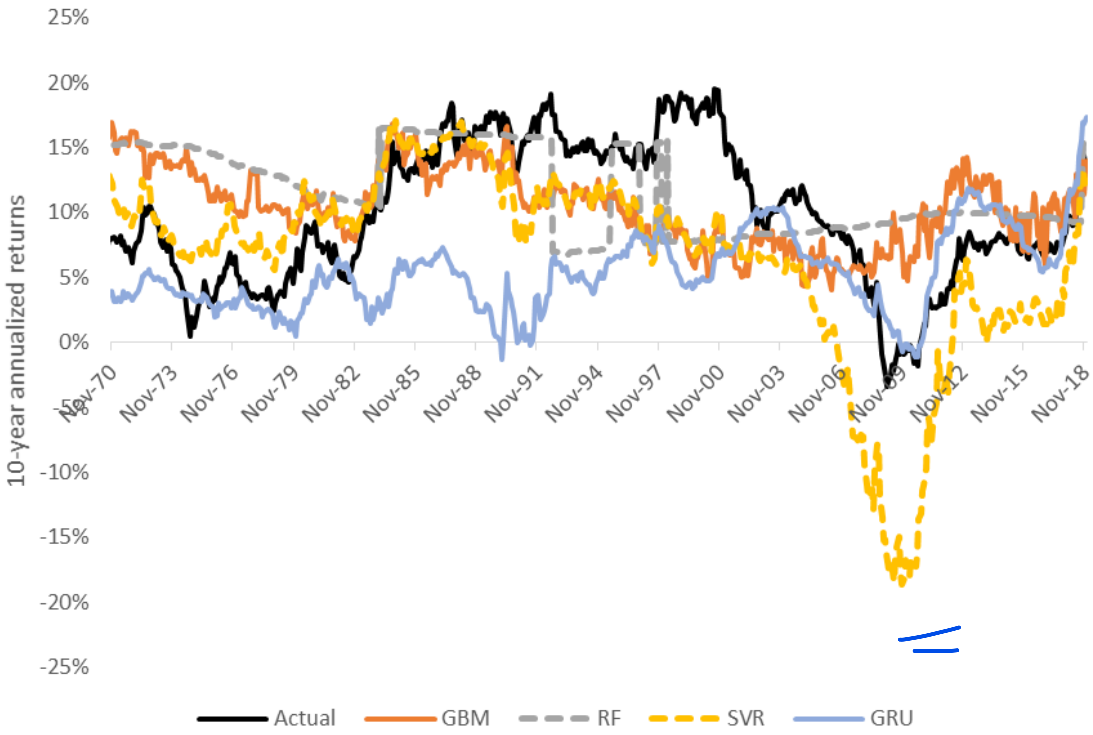
Note: We use data between January 1926 – December 1959 data as the training data and use data from 1960 to 2008 as the testing data

Source: Authors' calculations based on the data sources listed in the Appendix.

We use the training data, specified in Exhibit A1, to select the best hyperparameters for each model among all hyperparameter combinations shown in Exhibit A2. RMSEs are calculated based on the training data and the hyperparameters with the lowest RMSE for each ML method are chosen. We then recursively re-estimate the regression specification every month thereafter using the chosen hyperparameters to produce out-of-sample predictions.

For GRU we use the training data to select the hyperparameters that led to the lowest RMSE. We then apply the selected hyperparameters and the trained neural network to the testing data without the monthly recursive window. We make the change for GRU because the extensive computational power needed for building and testing the neural network makes it practically impossible to train GRU every month.

Exhibit A2: Prediction of GBM, RF, SVR and GRU under the Shiller approach



Notes: For the real-time analysis, the prediction (with the exception of GRU) are determined recursively at a monthly frequency, starting with January 1926 – December 1959 data. For each ML methods, we choose the hyperparameters with the lowest RMSE and re-estimating the regression coefficients every month thereafter using the chosen hyperparameters. For GRU, the prediction are determined by running the entire training data. The chosen hyperparameters with the lowest RMSE are then applied to the GRU for the entire period. The gap between the two lines represents forecast error.