# Machine Learning Classification Methods and Portfolio Allocation: An Examination of Market Efficiency[†]

Yang Bai
Kuntara Pukthuanthong

(Click for Updated Manuscript)
First Draft: May 1, 2020
This Draft: Jan 7, 2021

## Abstract

We frame the asset pricing problem as a machine learning classification problem[1]. The predictions on 3.34 million observations yield significant out-of-sample economic gains. Through directly measured accuracies, binomial tests suggest that the classifiers can extract forward-looking contents from historical information, implying imperfect information efficiency. The classifiers exploit the differences in return state transition uncertainties. As reflected by a pre-realization measure based on multi-class predicted probabilities, the classifiers are more confirmative in predicting high-trading-friction stocks. Consistently, only trading frictions contribute to out-of-sample predictability throughout 26,302 distinct stocks' lifetimes. The adjustment of the classifiers' favorance over certain return states increases the performance.

**Key Words**: Artificial neural network, big data, binomial test, classification, dropout additive regression tree, gradient boosting machine, information theory, machine learning, portfolio allocation, out-of-sample prediction, random forest, return state transition.

**JEL classification**: C14, C38, C55, G11, G14

[†] Yang Bai (Email: yangbai@mail.missouri.edu; Website: www.yangbai-finance.com) and Kuntara Pukthuanthong (Email: pukthuanthongk@missouri.edu; Website: www.kuntara.net) are from University of Missouri. This paper has benefited from the discussions with seminar and conference participants at University of Missouri, AFA 2021 Ph.D. Poster Session, Crowell Prize 2020 Seminar by PanAgora Asset Management, University of Miami Winter Research Conference on Machine Learning and Business 2021, and World Finance Conference 2021. We are particularly grateful for the insightful comments and suggestions from Ahmed Guecioueur (discussant), William Cong, Guofu Zhou, Massimiliano Caporin (discussant), and Alberto Rossi. This paper was also among the recipients of the Crowell Prize 2020 (third prize), and we thank the program committee at PanAgora Asset Management for their decision. We are currently updating the draft to address the comments and will release an updated draft soon. All errors are our own.

[1] Classification in machine learning means predictive categorization. A machine learning classification model is often called a classifier.

*Hal Weizman: "What is the efficient-markets hypothesis and how good a working model is it?"*

*Eugene Fama: "It's a very simple statement: prices reflect all available information. Testing that turns out to be more difficult, but it's a simple hypothesis."*

*Richard Thaler: "I like to distinguish two aspects of it. One is whether you can beat the market. The other is whether prices are correct."*

*— Are Market Efficient, Chicago Booth Review*, Jun 30, 2016

# 1 Introduction

Motivated by the success of machine learning in delivering out-of-sample (OOS) predictability, we introduce the new application of machine learning classification methods to one of the most widely studied problems in finance, measuring equity premium. Utilizing the novel framework, we study the drivers of the OOS performance of our machine learning methods and provide unique insights about market efficiency.

We frame the equity premium prediction problem as a single-label multi-class classification problem on return state transitions. Instead of modeling on future returns directly, our classification models (hereafter, classifiers) predict the probabilities associated with the future return states and the predicted probabilities decide the portfolio allocation. Our methods' significant economic and statistical performance relative to benchmarks adds another evidence about machine learning methods' capability to "beat" the market following the recent development in the financial machine learning literature (See Gu et al. 2020, Cong et al. 2020 and Chen et al. 2020, etc.).

Our framework provides unique benefits, as comparing to the numeric prediction methods, that allow us to provide unique insights about OOS predictability in financial machine learning and market efficiency. We provide the novel economic insights to the understanding of market efficiency through the introduction of the binomial test. Our tests show that the historical public information is not fully reflected in current prices, which leads to the performance of our classification-based portfolios (See also Cohen et al. 2020). We look into the return state transitions and attribute the performance to the classifiers' capability to exploit the different levels of uncertainty associated with return state transitions. Based on our empirical framework, we identify that the stocks with the highest trading frictions are the most predictable stocks in their lifetime. We also document that individual machine learning models can have their biased preference. Correcting the biased preference can dramatically increase the performance.

Specifically, we look at individual stocks with a data containing 3,342,486 monthly stock observations of 26,302 distinct common stocks listed on 3 major stock exchanges covering the period of 196301:201912. We first bucketize the stock returns with the cross-sectional deciles and split the returns into 10 return states. Using the historical information, including the individual stock returns with a lag of at least 1 month, the annual financial information with a lag of at least 6 months, the quarterly financial information with a lag of at least 4 months, the corporate event news with a lag of at least 1 month and the macroeconomic indicators with a lag of at least 1 month, we apply the classification methods to predictively classify the future return states of individual stocks and form portfolios based on the predictive classification. We show that our machine learning classification methods are powerful in portfolio allocation and our portfolios can produce huge OOS profits.

Unlike the numerical value approach, our approach allows us to produce the accuracy measure as a clear and concise metric of prediction quality directly. This helps deepen our understanding of the machine learning models. In contrast, the numerical value approach is limited by the modeling target and thus can only evaluate the prediction quality with error-based metrics. Moreover, since our final output for each stock at each time point is the probabilities associated with the potential outcome return states, this leaves a room for us to study the models' subjective feelings, including modeling certainty and modeling confidence. Unlike the numerical value approach, we can form pre-realization measures before looking at the true outcome in the future period and study the models' certainty level and confidence level about the predictions during the prediction process before the looking at the ground truth.

Next, taking the advantage of the unique features of our framework, we study the drivers behind OOS predictability and market efficiency from information efficiency viewpoint. The introduction of the classification methods provides unique benefits for us in both the modeling process and the evaluation process. In the training process, we take advantage of a clear relation between the classification methods and the information theory. We measure the quality of information extracted from the predictors with cross-entropy and train our models with the optimization goal of reducing information uncertainty. The classification methods also allow us to directly measure model performance through accuracy calculated as the correct proportion of predictive classification. Measuring prediction performance through accuracy not only is easy and explicit but allows us to conduct formal statistical tests. We further introduce the binomial test to compare our prediction accuracies against the no information accuracy. In addition to the no information accuracy, we also look at the benchmark implied by the assumption that the stock return follows a memoryless process, and the best prediction of the future return state is today's return state. Statistically, our prediction accuracy dominates both of the benchmarks.

The no information accuracy is the highest accuracy that a classifier with no information or limited information can provide. In the language of machine learning, it is the accuracy delivered by a naive classifier, which labels the return state of each observation in our sample with the most populated return state in the sample. Based on the rational investor assumption, which leads to the efficient market hypothesis, investors have consistent beliefs and enough information about the distribution of macroeconomic variables (Sargent 1994, Barberis and Thaler 2003). If the market is efficient and all information is reflected by prices, investors know about the return distribution but are not able to predict the future beyond the distribution of the returns (See also, Fama 1970, 1991, and 1998). The no information accuracy is thus a theory implied benchmark to evaluate whether there exists information about the relation, as captured by a model, between future return states and historical information.

Through a set of binomial tests against the no information accuracy, we trace the good performance of our classification portfolios to the statistically significant prediction accuracies. This further implies that our classifiers provide meaningful information about the relation between future return states and historical information. Future return states are thus conditional on historical information and the predictability is established. In other words, historical information is not fully reflected by the current prices and the market adjusts the prices in the next period following the direction predicted by our models.

Our findings implies that investors applying modeling methods to form portfolios in a way similar to ours can make profits systematically higher than what the market can offer with better risk-return tradeoffs. Because of the existence of information that has not been incorporated into prices, sophisticated investors, for example, Renaissance Technologies, can extract useful information about market prices through complex analytical tools. The generated information may not be available to the public and thus may create information asymmetry that sophisticated investors can benefit from. The investors who devote resources to obtain information are thus compensated by the market like what is demonstrated with our OOS performance. The question is how many people can access the data, the sophisticated analytic tools and the tradability that comes along with the strategy. If too many do, then the signals generated will be easily negated by crowding. Our setup with the models making OOS predictions for the period from 1992 to 2019 with the in-sample (IS) training period from 1963 to 1991 shows that the rules the models learned are still functional in the OOS period, which implies that the machine learning classification methods have not been overexplored in recent decades. The fact that past returns and past corporate announcements contribute to the OOS predictability also questions the weak-form and the semi-strong form of market efficiency.

Our unique setup allows us to examine the drivers behind the predictability closely and we provide unique economic insights about the source of predictability and the insights about classifiers. We document that there exists a substantial imbalance in the return state transition process. The transitions related to extreme return states are with higher certainty indicating lower market efficiency, which question the role that the market segments of extreme return states play in market efficiency. At the same time, through directly measure OOS prediction accuracy, we report that the individual stocks with higher trading frictions throughout their lifetime in our sample are associated with higher OOS predictability. We construct a pre-realization modeling certainty measure and show that models feel more certain when they are making predictions on the stocks with higher trading frictions. In the end, we show that the machine learning models can have systematically biased preferences over certain outcomes, which can decrease the performance of the models (See also Fuster, Goldsmith-Pinkham 2020 and Ramadorai and Walther 2020).

## 1.1 Contribution

In summary, we contribute to multiple branches of the literature, including the return predictability literature, the financial machine learning literature and the market efficiency literature. Our contribution is eightfold.

Specifically, first, our introduction of classification methods completes the methodological picture of the study on future return explanation in financial machine learning (as in Gu et al. 2020). We categorize the returns into return states and predictively classify the stocks into different possible future return states with historical public information. We demonstrate 2 machine learning architectures, 4 types of algorithms, and 22 models. We include shallow neural networks, deep neural networks, random forests, dropout additive regression trees, and stochastic gradient boosted trees.

Our approach advances the extant machine learning techniques that is applied in asset pricing literature such as Gu et al (2020), Feng et al (2020), and Cong et al (2020). Despite using less training data and including only the stocks listed on 3 major exchanges, the performance of our portfolios is competitive and on par with the performance reported in the literature with the numeric prediction methods. In the OOS comparisons, the portfolios based on the predictions of the classification models can generate average returns, the volatility of the returns, skewness of the returns, Sharpe Ratios (SR), certainty equivalent returns (CEQ), and maximum drawdowns (Max DD) that are better than those based on the market.

To illustrate, the Sharpe ratio (SR) of 0.87 delivered by our best equal-weight model is higher than the SR of 0.707 by the best equal-weight portfolio reported by Gu et al. (2020). The SR of 0.42 delivered by our best value-weight portfolio is higher than the SR of 0.38 delivered by the best value-weight portfolio reported by Gu et al. (2020). The market portfolio delivers an out-of-sample (OOS) SR of 0.13 and an OOS SR of 0.12 for the two corresponding weighting schemes during the same time[2].

The good performance of our portfolios is not from neither taking high leverage nor the concentration of portfolio weights in the microcap stocks. None of our portfolios requires leverage beyond the relaxation of short selling constraints. When we eliminate the bottom 5% and 10% capitalization stocks, the performance of our portfolios does not disappear. After we apply adjustments to the classification mechanism, even under an extremely conservative situation where we exclude the stocks with the size below the 50% level of the sample, our models can provide significantly higher SRs comparing to what the market can provide (See Section 4.).

Second, we introduce accuracy as a performance metric under the classification framework and the adoption of the binomial test contribute to the predictability literature and expands the toolbox for empirical asset pricing. The direct measurement of the accuracy as the correct proportion of predictions is only available to classification problems. In numeric value predictions, all metrics are

[2] Note that our SRs are not annualized nor adjusted with R-squared. Either adjustment can greatly magnify SR. To annualize SRs, one will just multiply SRs by $\sqrt{12}$.

based on prediction errors and are hard to directly measure the accuracy of predictions. We trace the good performance of our classification portfolios to the accuracy of the return state predictions. We also propose to use prediction accuracy as a metric to study information efficiency in financial markets. We carefully analyze the in-sample (IS) and the OOS prediction accuracies and provide an explanation of the good portfolio performance from the angle of information theory.

Third, we introduce the binomial test to the asset pricing literature, which enables us to conduct a meaningful statistical test on the prediction accuracy. Along with the binomial test, we also introduce the no information accuracy. Across multiple setups, we show that our models deliver statistically meaningful predictability and are time-invariantly applicable to generate predictions of future return states.

The binomial tests on prediction accuracies against the no information accuracy have profound meaning to the study of market efficiency. The no information accuracy is an accuracy under the assumption of the efficient market. In other words, the no information accuracy is the highest accuracy of prediction assuming that no further information can be generated to describe the relation between future prices and historical information. Therefore, a binomial test on the prediction accuracy against the no information accuracy is not only a test on the predictability but also a test of the market efficiency. The statistical significance found in our binomial tests against the no information accuracy implies the generation of information through our models about future return states based on historical observations. This piece of information is not reflected by the current prices and therefore not shared by most of the market participates. At the same time, this piece of information does generate OOS profitability. This naturally questions the correctness of the prices, i.e., the prices may not be correct as people can make a profit with public information.

Across the entire CRSP-COMPUSTAT sample, the results suggest systematic trading opportunities based on historical information to generate monthly excess profits. The profitability generated by trading on the information from complex tools has sizable economic gain that is difficult to ignore. It is worth noting that this price related to the generated information should have been eliminated by arbitrage. As such, the current market prices may not be correct or may not fully reflect all public information. This also provides an answer to Thaler's question about whether the prices are correct.

The generation of the new information also has an important implication that is directly related to the strong-form market efficiency. The information generation implies that sophisticated investors, such as Renaissance Technologies, which has access to humongous amount of information and complex tools similarly to the input used by machine learning classification methods, can generate information about future return states from historical observations. They can apply the new information to their trading. The generated information, depending on the analytical tools, is likely to remain unique and monopolistic. In other words, the private information may not need to be insider information that is known to the management team of a firm but can be generated based on analyzing historical information. Our results confirm that there is a possibility for sophisticated investors to manually introduce information asymmetry to the market.

Taken together, we are the first to introduce the binomial test and the accuracy metric in the study of return predictability and market efficiency. By far, the binomial test on the prediction accuracy that indicates the new information generation is also a unique contribution to the finance machine learning literature and the market efficiency literature. The generation of new information also helps us explain the good performance of successful portfolio allocation strategies from an information point of view.

Our findings on market efficiency are consistent with the microstructure literature which shows theoretically that the information efficiency is conditional, and a full informationally efficient market is impossible. Extra rents can only be earned on genuine information not available to all (Grossman and Stiglitz 1980). Our findings on market efficiency are also consistent with recent literature indicating that prices are lazy, and information may be included in the prices with lags (Cohen, Malloy and Nguyen 2020).

Fourth, researchers have shown evidence of increasing market efficiency and shocks that are associated market efficiency (See Linnainmaa and Roberts, 2018; Rösch, Subrahmanyam, and van Dijk, 2017). We investigate further and identify which stocks and associated characteristics that tend to violate market efficiency. Utilizing the unique setup and the directly measured OOS accuracy, we look into the lifetime of 26,000 distinct individual stocks. We identify what stocks are more predictable through regression analysis. We group the effects from the stock characteristics and report that the stocks with higher trading frictions are more predictable during their lifetime. Our results indicate that the trading frictions are the only group of characteristics that are positively related to OOS predictability. Other groups of characteristics, including momentum, are negatively related to the OOS predictability. This finding is critical, since the predictability is a signal for the existence of historical information that has not been fully incorporated into prices. At the individual stock level throughout the stocks' lifetime, our findings of the trading frictions further confirm that different stocks may experience different levels of market efficiency.

Fifth, by introducing the novel empirical framework to study the return state transition, we provide a unique set of new economic insights. We directly measure return states, we can compare the easiness of predicting different return state transitions. For the first time in the literature, we demonstrate through the true return state transition probability matrix that the return state transitions are not uniformly distributed. The center of the true transition probability matrix is distributed more uniformly, which indicates a higher level of uncertainty. The corners of the true transition probability matrix are with the highest transition probabilities indicating a lower level of uncertainty. The different levels of uncertainty question the role that the market segments of extreme return states play in market efficiency. Higher uncertainty is related to lower predictability, indicating a higher level of market efficiency and vice versa. Meanwhile, we show that our models benefit the most from the most certain transitions and almost give up the more uncertain transitions, revealing the driver of the predictability. We are among the first to supply the asset pricing literature with unique economic insight behind the success of the machine learning portfolio allocation.

Sixth, we introduce a modeling certainty measure defined as the variance across all the predicted probabilities associated with the potential outcomes. Ideally, if the model feels good about the prediction, the variance across all the predicted probabilities will be high, indicating that the model is more certain about what can happen in the next period and that the model can distinguish the potential outcomes associated with high probabilities from those associated with low probabilities. Following similar logic of our regression analysis at the individual stock level for the predictability, we examine the model feeling about individual stocks throughout their lifetime. With our modeling certainty measure, we report that the stocks with higher trading frictions can make the models feel better about their predictions. Note that this measure can be calculated prior to the realization of the ground truth outcome in the next period. This issue is important as it indicates our measure is not subject to look ahead bias.

Seventh, taking the advantage of our response variable, we can closely study the biased preference of our models. We show that a classifier can concentrate its prediction on a small group of potential outcomes, and this can lead to systematically biased preferences. The biased preferences will reduce the OOS performance of our method or lead to a complete failure of the OOS performance. We further report that after adjusting the classification mechanism based on the predicted probabilities, the OOS performance increases systematically across all our models. The failed model in our default setup, ANN3 128, also delivers significant OOS performance after the adjustment of the default classification mechanism.

Finally, through our demonstration of the classification methods, we analyze the training process. We contribute to the identification of the important predictors of return states. We construct cross-sectional tests in the spirit of Fama and French (2018) by splitting the CRSP sample into odd number months and even number months. The cross-sectional OOS tests show that our models have good CS OOS explanatory power. Besides, we look into the training process of the CS models and the time series (TS) models. In the training process of both the CS models and the TS models, the industry information, the corporate announcements, the macroeconomic indicators, and the historical return information all make an important contribution.

As all our predictors are lagged by at least 1 month and many of the firm characteristics are lagged by at least 6 months, these findings are interesting to the study of market efficiency, especially considering that we do not update our models in the TS OOS testing periods with at least a time length that is close to 30 years[3]. Coupled with the OOS portfolio performance based on the model predictions, we conclude that the historical public information, including the past returns and the historical corporate announcements, can help predict future return states.

[3] We are looking into dissecting the insights for weak form and strong form market efficiency by separately developing models that use only past trading information and models that use only past corporate news. Results will be included in the next update of the draft.

## 1.2 Literature[4]

We review the recent development in the finance machine learning literature below. Our review is by no means an exhaustive list of the works in the finance machine learning literature and we try to include the works that are the closest to ours in a chronological order (as of the draft date) based on the publication date for the published papers and the latest update date for the working papers. We categorize the papers based on the model implementations.

### 1.2.1 Characteristics

Brandt, Santa-Clara and Valkanov (2007) is a pioneer in leveraging characteristics in the portfolio allocation problem. They parameterize the portfolio weight of each stock as a function of the stock's characteristics and they estimate the weights of the stocks included in the portfolio with the maximization of the representative investor's utility. The optimal portfolio relative to holding the market provides an in-sample (IS) CEQ gain of 11.1% and 5.4% OOS CEQ gain. Green, Hand and Zhang (2016) is the pioneer of studying the large number of firm characteristics. They construct a sample of more than 100 firm characteristics based on stock performance and financial information. They show that there are 12 characteristics that are reliably independent in contributing to the return predictability during their sample period and the predictability drops after 2003.

### 1.2.2 Tree Models

Moritz and Zimmermann (2016) introduce the regression trees in the cross-sectional pricing. Using the regression trees they show that the past short-term returns are the most important predictors for the future returns. They sort the portfolios with tree structure. They show that the conditional portfolio sorts through tree structure improve predictions significantly over Fama-MacBeth regression. Rossi (2018), using boosted trees, documents that the non-linearity of the popular Goyal and Welch (2008) predictors can time the market. He emphasizes that the relation between predictors and the best allocation to risky portfolios is non-linear. Bryzgalova, Pelger and Zhu (2020) demonstrate the advantages of applying pruning in the selection of the sorting methods to improve the empirical asset pricing models.

### 1.2.3 Neuron Networks

Chen, Pelger and Zhu (2020) focus on the neuron network models and asset pricing. They combine 3 neuron networks and essentially generalize the linear pricing kernel under the framework of neuron networks. They introduce the generative adversarial neuron network models to the playground of fine

[4] We are updating the literature review to incorporate more recent papers that are circulating on the topics of finance machine learning. This section will be updated soon. The last update was in July, 2020.

search of the best SDF by identifying the assets that are hardest to model. They also enforce the non-arbitrage constraint to the loss function in the architecture of the networks. Aubry, Kraussl, Manso and Spaenjers (2020) introduce the machine learning methods to the playground of illiquid assets. Specifically, they apply neuron networks to a data with one million painting auctions based on visual and non-visual characteristics of the art pieces. They show that their methods perform drastically better than the traditional pricing methods. Feng et al. (2019) propose the use of neuron networks in the extraction of hidden features and augment the hidden features in the pricing models.

### 1.2.4 Tree Models and Neuron Networks

Gu et al. (2020) demonstrate the powerful pricing capability of the neuron network models and the tree models. Their experiments show the possibility to double the performance of leading regression-based strategies. They also try to form OOS portfolios by predicting the stock returns first and then forming portfolios the stocks based on predicted return. Their best equal-weight strategy coming from a 4 hidden layer neuron network delivers a shockingly 27.1% return on annualized basis. In a concurrent work of ours, Wolff and Echterling (2020) construct 37 stock characteristics and also characterize the portfolio allocation problem as classification problem. They apply neuron networks and tree models to S&P 500 constituents with 21 years of weekly data. They show that their models can also be applied to STOXX Europe 600. Bianchi, Buchner and Tamoni (2020) apply machine learning methods including neuron networks and tree models to the bond market. They demonstrate the superior performance of the machine learning methods in predicting bond returns.

### 1.2.5 Related Works

The two closest related works to our paper are Gu et al. (2020) and Wolff and Echterling (2020). Our paper is distant in many aspects from Gu et al. (2020). First, our introduction of classification is fundamentally different from the implementation of Gu et al. (2020). In fact, we view the portfolio allocation as a selection problem of the stocks, while Gu et al. (2020) view the portfolio allocation as an estimation problem of the stock returns. In other words, we model on the probabilities of return state transitions conditional on historical information and Gu et al. (2020) model on the numeric value of returns. Taking the neuron network models as an example, our neuron network models all include a soft-max output layer of 10 neurons that gives us the probabilities of a stock being in one of the 10 return states in one period ahead, while neuron network models in Gu et al. (2020) all have a linear output layer of 1 neuron and output a return prediction. Similarly, our boosted tree models are based on multi-class probabilities, while the boosted tree models in Gu et al. (2020) are based on linear regressions. Our output gives a better sense of relative performance of stock returns and the probability of occurrence.

Our paper is also very different from Wolff and Echterling (2020). First, we model on monthly returns covering 196301:201912 including all 26302 stocks out of all 33004 securities. We include

332 predictors covering historical returns, firm characteristics and macro indicators, while Wolff and Echterling (2020) include 37 predictors and 21 years of weekly data (199901:201912). In addition, to the specific purpose of our study, we characterize the returns into 10 return states independent of market return, while Wolff and Echterling use binary categorization with reference to market return.

The most important aspect that distinguishes our paper from Gu et al. (2020) and Wolff and Echterling (2020) is the scope of the studies. We choose machine learning classification methods specifically because of their relation with the information theory and the testing metrics that can be adopted. Beyond the modeling aspects and the predictive power, we attempt to provide additional insights about market efficiency. We also aim at providing new economic intuition about why the machine learning methods can produce portfolios that outperform the market. Through our comprehensive analysis and the demonstration of our 22 models, we show explicitly that the investors can generate new information and the market is unbalanced in terms of the transition probabilities.

The remainder of the paper proceeds as follows. We specify the models, metrics, and the empirical setup in Section 2. In Section 3, we demonstrate the OOS performance of our classification based portfolios with economic and statistical metrics. We analyze the performance through accuracy and discuss binomial tests. We also discuss unique economic insights about the portfolio allocation strategies, predictable stocks, and machine learning models in Section 4. In Section 5, we document the cross-sectional explanatory power and predictor contribution. We conclude in Section 6.

# 2 Methodology

We provide a general description of our methods in this section. We explain the basics of our modeling process and tests, including our model specifications, validation and hyperparameter tuning, testing metrics and sample splitting. We provide details so that readers with limited experience

## 2.1 Model Training and Validation

### 2.1.1 A Brief Introduction to Classification and Information Theory

A classification problem is a choice making problem. For example, given a picture capturing an animal with 2 possible outcomes, cat and dog, a classification problem can be framed as the question: is the animal in the picture a cat? This is a binary choice question. If the answer is yes, we know that the picture captures a cat. If the answer is no, the picture captures a dog. This is a typical binary classification problem with one class being the cat pictures and the other class being the dog pictures. The task in this classification problem is to find a strategy to label a picture to be either a picture of a cat or a picture of a dog. A strategy is referred to as a classifier or a model in machine learning literature. If we have a classifier that always guesses that the animal captured by any picture is a cat, then this classifier is a naive binary classifier. The classification outcome of a given picture, i.e. cat or dog, is called the label of the picture. A classification problem is not limited to have 2 candidate outcomes nor a single label. For example, the question that asks “what is the weather tomorrow?” is a multi-label multi-class classification problem. Specifically, for example, an answer to the question can include 2 labels, one about the weather condition and the other about temperature. The candidate outcome weather conditions can include rainy, snowy, sunny, etc. The candidate outcome temperature can include 3 levels: hot, mild, and cold. In this paper, we frame our portfolio allocation practice as a single label multi-class classification problem.

In Table 1, we demonstrate how we frame a portfolio allocation problem as a classification problem. We cross-sectionally rank individual stock returns by trading month, put them into their corresponding deciles and use the deciles as the classes of return states. For example, if a stock falls into the lowest decile in a trading month, we define the true label of the stock as the class of return state 1. A stock in return state 1 means that the stock delivers a return that is among the worst-performing returns of the trading month. A stock in return state 10 indicates that the stock is among the stocks delivering the best performing returns of the trading month. In later sections, we refer to the return states with the numbers specified in Table 1. In short, small number return states indicate bad performing return states while large number return states indicate good performing states. Note that we make the lower bound the inclusive bound; therefore, we have a slightly unbalanced 10 classes.

In Section 6 of Claude Shannon’s (1948) seminal paper, Shannon introduces the concept of a bit as the unit for information and the famous Shannon entropy, or information entropy (entropy hereafter), as the measure of the average level of information, or the amount of randomness, in the unit of bits. An arbitrary parent choice question, for example, can be decomposed into a series of binary choice sub-questions and the entropy summarizes the average number of the binary choice sub-questions to answer such that the parent choice question can be answered. Higher entropy means that there is more uncertainty. The entropy can be defined as

$$Entropy(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-,$$

where $S$ is a Bernoulli trial with 2 possible outcomes $\{+,-\}$ that are mutually exclusive and $p_+$and $p_- = 1 - p_+$ represent the probabilities of the two possible outcomes respectively. When the entropy is seen as an uncertainty measure of the information, the greater the entropy is the greater the uncertainty is. In a binary case, the entropy is at its largest when $p_+ = 1 - p_- = 0.5$. In other words, a Bernoulli distribution that is close to a binary outcome discrete uniform distribution is with the highest uncertainty and is more likely to yield a surprising outcome. When a Bernoulli distribution departs from the binary outcome discrete uniform distribution, the uncertainty decreases, and a random draw from the Bernoulli distribution is more likely to deliver an unsurprising outcome. Specifically, for example, consider the weather outcomes of snow versus not snow during winter. Florida has a lot lower uncertainty of snow comparing to the uncertainty level of snow in Missouri as the probability distribution of snow versus not snow in Florida is skewed towards snow and concentrated in not snow while the probability distribution of snow is closer to a uniform distribution in Missouri.

A generalization of entropy to compare the difference between two probability distributions yields the cross-entropy. Given a parent choice question and our best strategy to form binary choice sub-questions, the cross-entropy measures the average number of binary choice sub-questions that we need to answer with our best strategy in the environment with the true probability. In the training process of a machine learning model, we have the observed probability distribution and we will select the best strategy, by comparing the candidate estimates of the distributions, to form the binary choice sub-questions and answer the parent choice question. The best strategy is associated with the lowest cross-entropy and thus also with the lowest information uncertainty. The information uncertainty is associated with the possibility of information loss. In other words, the cross-entropy can be thought of as a measure to compare the observed probability distribution and the predicted probability distribution.

### 2.1.2 Loss Function and Optimization

In a training process of multi-class classification problems, we want to minimize the overall errors and balance the by-class performance of the classifier. Popular choices of loss functions include accuracy, information gain based on entropy, mean error across the classes, etc. We adopt cross-entropy as the loss function to minimize the information uncertainty between our predicted distribution and the observed distribution. Specifically, we achieve this by comparing our predicted distribution against the true distribution with cross-entropy. The formal definition of cross-entropy is:

$$\begin{aligned} Cross\ Entropy =& \mathbb{E}_{p_i} \log_2 \frac{1}{\hat{p}_i} \\ =& \sum_i p_i \log_2 \frac{1}{\hat{p}_i}, \end{aligned}$$

where $p_i$ is the observed probability of outcome $i$ and $\hat{p}_i$ is the predicted probability. Higher cross-entropy represents higher information uncertainty associated with the use of the predicted probability to approximate the observed probability. In other words, higher cross-entropy represents lower information decoding quality with the predicted probability compared to the observed probability. When we train a classification model, for each iteration of weight update, we aim at minimizing the cross-entropy through the adjustment in model weights. When we conduct hyperparameter training in the format of the grid search, we also select the best model based on cross-entropy. By using cross-entropy as the criterion to adjust weights and select hyperparameters, we can obtain a model that reduces the uncertainty of the information extraction.

### 2.1.3 Model Specification, Validation and Hyperparameter Tuning

We describe the models in our consideration inTable 2. We include 2 architectures covering shallow neuron networks, deep neuron networks, dropout additive regression trees, random forest, and gradient boosting machine, a total of 22 separate models. In our specification and model training, to understand the effect of the major architectural parameters, which control the model complexity, we do not add the architectural parameters, such as the number of hidden layers in a neuron network or the number of maximum depth in a tree model, into the hyperparameter tuning process. Through training and evaluating models with the different key parameter specification, we can see a clear trend later that the differences in the key structural parameters have a strong influence on the performance of the models.

Specifically, we include neuron networks with 1 to 4 hidden layers. Our tree models are with maximum depths of 2 to 8. The number of hidden layers controls the complexity of interaction across predictors. The number of maximum depth limits the maximum number of leaves that a tree can grow. The complexity of computation can increase exponentially as we increase the depth of the tree models. The last column in Panel A of Table 2 presents the structural capacity of our models. The numbers in curly brackets correspond to the number of neurons in the specific hidden layer. For example, the model ANN4 128 has 128 neurons in the first hidden layer, 64 neurons in the second layer, 32 neurons in the third hidden layer, and 16 neurons in the fourth hidden layer. Thus, the numbers of neurons in the curly brackets for ANN4 128 is $\{128,64,32,16\}$. For our DART models, beyond the important parameters summarized in Panel A, we also specify the dropout rate as 10%. This dropout rate can help generalize the model. To save on computation resources, we also apply an early stopping mechanism to all of our models. If a model in the training process does not improve the loss function by at least 0.00001 for 3 consecutive rounds, the training stops.

Panel B in Table 2 presents the additional specification information that applies only to neuron network models. Relu is the popular choice of activation function for the hidden layers in the recent finance machine learning literature. However, we did not use Relu as the hidden layer activation function. We want to avoid the dead neurons in the deeper layers of deeper networks. Our selection of Tanh function is famous for its robustness. Because our interest of the study is the return states as classes, we specify the output layer with SoftMax function, which transforms the inputs from the last hidden layer to the probabilities. We set the output layer to include 10 neurons, corresponding to 10 possible return states. For each instance fed to our neuron networks, each neuron in the output layer will produce a probability representing the likelihood that the instance belongs to the associated return state. In the end, we categorize a stock to one of the return states associated with the highest probability.

In any of our neuron networks, we have 3 layers of transformations starting from the input layer. Consider a neuron network with 1 input layer, 1 hidden layer of 1 neuron and 1 output layer with 10 neurons. Let us denote the input layer as $X$. We form transformation through activation function taking linear combination of the input layer as its input. The transformation is defined as $Z = \sigma(\alpha_0 + \alpha^T X)$, where $\sigma$ is the Tanh function. Then, we further transform the output of the Tanh function through another linear combination and connect linear combination with the output layer of SoftMax function. Specifically, we first collect the linear combination taking $Z$ as input and denote the linear combination as $T_k = \beta_{0k} + \beta_k Z$, where $k$ is the number of neurons in the output layer. Then, we connect $T$'s to the observations through the SoftMax function $g(T) = \frac{e^{T_k}}{\sum_{l=1}^{10} e^{T_l}}$. During training, we adjust the weights in all layers and bring the SoftMax function to produce probabilities for individual observations as close to the real probabilities as possible. In a tree model, the logic is similar but different. In our tree models, the training process is done with the sub-sample of individual classes. Each tree will select a class of return state and use the subsample of the training set containing the selected return state to learn and adjust the weights. The weights of different trees are not directly interacting with each other until the summarizing step when the model pulls all the information about individual classes together and conducts a majority vote process to decide the prediction. When the majority vote step takes place, the probabilities of individual classes will be summarized and scaled to reflect the overall probabilities with a summation of 1.

Panel C of Table 2 presents the hyperparameters that we want to tune with our validation strategy. As we separate the architectural parameters from the hyperparameter set to help us understand more about the influence of model complexity, we only have a limited number of hyperparameters to tune with our model. Specifically, for neuron networks, we tune the L1 regularization parameter which decides the penalty put on the weights similar to the regularization in the lasso regression. Our neuron networks prefer the finite L1 regularization. We tune the sampling rate for training data and the sampling rate for the predictors as a control for the generalization of the tree models. We take cross-validation as the validation strategy for hyperparameter tuning. We choose cross-validation, instead of constructing a separate validation sample as being implemented by Chen et al. (2020) and Gu et al (2020), to take the advantage of the data coverage and avoid the loss of OOS testing observations. Specifically, we separate the training data set into 5 subsamples in chronological order and conduct 5-fold cross-validation.

## 2.2 Performance Evaluation

To better communicate our empirical findings, we describe the metrics that we refer to. For model-based portfolio allocations, it is important for us to understand both the economic performance and the statistical performance. Therefore, we list out both the economic metrics and the statistical metrics.

### 2.2.1 Economic Metrics

The purpose to evaluate a model-based portfolio economically is to understand whether the portfolio is successful in terms of commonly used traditional measures. Specifically, we refer to Sharpe Ratio (SR) and Certainty Equivalent Return (CEQ) in the evaluation of the risk-return trade-off. Portfolios with better performance in terms of risk-return trade-off have higher SR and CEQ. We define SR as

$$SR = \frac{\mathbb{E}(R - R_f)}{\sigma(R - R_f)},$$

where $R$ is the return generated from a portfolio of interest and $R_f$ is the risk free rate of return. For the long-short portfolios, we define the SR as

$$SR_{long-short} = \frac{\mathbb{E}(R_{long} - R_{short})}{\sigma(R_{long} - R_{short})},$$

where $R_{long}$ is the return generated from holding the long position of of the predicted good performing stocks and $R_{short}$ is the return generated from holding the long position of the predicted bad performing stocks. We define the long-short SR in this way as the long-short portfolio is a theoretically zero investment portfolio. Following DeMiguel, Garlappi and Uppal (2009), we define CEQ as

$$\widehat{CEQ}_k = \hat{\mu}_k - \frac{\gamma}{2}\hat{\sigma}_k^2,$$

where $\hat{\mu}_k$ is the estimated mean of the return from the asset $k$ and $\hat{\sigma}_k^2$ is the variance of the return. $\gamma$ in the above expression stands for the risk aversion coefficient and we specify $\gamma = 1$ following DeMiguel et al (2009) and Goyal and Welch (2008).

In addition to these most popular economic metrics, we also provide basic metrics to evaluate the profit and loss. We specify the cumulative return as

$$Y_{t:t+n} = \prod_{i=t}^{t+n}(1 + R_i) - 1$$

, where $R_i$ is the return from the portfolio of interest in the month $i$ and $n$ stands for the number of periods in the investment window. Our cumulative return is therefore defined as the product of gross return net of the initial investment cost. We take the notation of our cumulative return and include maximum drawdown in our evaluation defined as the following:

$$MaxDD_{t:t+n} = \min_{t:t+n}\{\frac{Y_{i+1} - Y_i^{peak}}{Y_i^{peak}}\},$$

where $i$ is a trading month during the investment window $t : t+n$ and $Y_i^{peak}$ is the highest cummulative return until time point $i$. Finally, following Gu et al. (2020), we provide turnover defined as

$$Turnover = \frac{1}{n}\sum_{i=t}^{t+n}\left(\sum_{j}\left|w_{j,i+1} - \frac{w_{j,i}(1+r_{j,i+1})}{\sum_k w_{k,i}(1+r_{k,i+1})}\right|\right),$$

where $w_{j,i}$ represents the weight of stock $j$ during month $i$ in a portfolio.

### 2.2.2 Statistical Metrics

For our classification models, we introduce and report a range of metrics that focuses on model accuracy from both the angles of overall accuracy and balance of the prediction accuracy across different classes. To better introduce the statistical metrics, suppose that we have a binary classification problem and a classifier making predictions. Consider the following matrix which compares the true values and the predicted values:

| | Reference Positive | Reference Negative |
|---|---|---|
| Predicted Positive | A | B |
| Predicted Negative | C | D |

, where the rows indicate the predicted value and the columns are the references of ground truth. The letters A, B, C, and D stand for the number of observations. Specifically, for example, A stands for the number of the observations with the true positive label that also are predicted to have a positive label. In such a case, A is the number of correctly predicted positive observations or the true positives. Similarly, D is the number of correctly predicted negative observations or the true negatives. B and C stand for false positives and false negatives. A matrix that compares the number of predicted observations with the ground truth is called a confusion matrix.

With the basics of the confusion matrix being introduced, we can further introduce the popular metrics that evaluate the performance of a classification model. First, referring back to the confusion matrix example above, we define sensitivity and specificity as

$$\begin{aligned} Sensitivity &= \frac{A}{A+C} \\ Specificity &= \frac{D}{B+D}. \end{aligned}$$

Sensitivity measures the accuracy of the predicted positives, while Specificity measure the accuracy of the predicted negatives.

In an ideal situation for a binary classification problem, we want to maximize the overall accuracy or the number of A+D and at the same time keep a balance between making positive and negative predictions. A classic example can be the detection of cancer. The proportion of cancer patients over the entire population who take the cancer screening is a relatively small number. Therefore, in such a situation, a classifier can gain a very high accuracy if the classifier just simply predictively label all people in the cancer screening as negative. However, the classifier will then fail to detect any potential cancer patients. Another example is information. For an extremely skewed distribution such as the chance of raining in the Sahara desert, telling the information receiver that the rare event will happen, ex., it is going to rain in the Sahara desert, carries more information comparing to telling the information receiver that the outcome associated with the largest possibility is likely to happen, ex., it is not going to rain in Sahara desert. When we train a machine learning model, the overall accuracy is just one aspect that we care about. We also care about whether the model generates new information and has a meaningful detection rate for each class. Thus, balancing the true positives and the true negatives is important.

Similar to sensitivity and specificity, we can have prevalence and detection prevalence. The prevalence measures the ground truth percentage of the sample being positive and the detection prevalence measures the predicted percentage of the sample being positive. Specifically,

$$\begin{aligned}
Prevalence =& \frac{A+C}{A+B+C+D} \\
Detection\ Prevalence =& \frac{A+B}{A+B+C+D}.
\end{aligned}$$

Beyond the above metrics that look into individual aspects of the predictions, there are 3 comprehensive metrics: the F1 score, the balanced accuracy and Cohen's Kappa:

$$\begin{aligned}
F1 =& \frac{(1+\beta^2) \times Precision \times Recall}{(\beta^2 \times Precision) + Recall} \\
Balanced\ Accuracy =& \frac{Sensitivity + Specificity}{2}, \\
\kappa =& \frac{p_o - p_e}{1 - p_e}
\end{aligned}$$

where $\beta$ in the F1 score is the type II error. The precision and the recall in F1 score are defined as $Precision = \frac{A}{A+B}$ and $Recall = \frac{A}{A+C}$. $p_o = \frac{A+D}{A+B+C+D}$ in $\kappa$ is the relative agreement observed between the ground truth and the prediction, while $p_e = p_+ + p_-$ measures the probability that the agreement between the prediction and the ground truth is random, where $p_+ = \frac{A+B}{A+B+C+D} \cdot \frac{A+C}{A+B+C+D}$ and $p_- = \frac{C+D}{A+B+C+D} \cdot \frac{B+D}{A+B+C+D}$. Note that F1 does not take into account the true negatives and thus have limitation on evaluating the results with consideration on the balance between true positives and the true negatives.

### 2.2.3 Accuracy and Binomial Test: A Novel Empirical Framework

In addition to the above metrics, we also introduce accuracy as a direct statistical metric of prediction performance. For any model, Accuracy is defined as the proportion of correct predictions. Note that this Accuracy is different from Balanced Accuracy. We also introduce a formal test on the statistical significance of the prediction accuracy. For any classification problem, after the classifier makes the prediction, we have two types of observations, the correctly classified observations and the incorrectly classified observations. Therefore, the prediction for an observation is a Bernoulli trial for a classifier. The number of correctly classified observations can be seen as the number of successes and the number of incorrectly classified observations can be seen as the number of failures. Following this logic, the prediction accuracy measured as the number of correctly classified observations over the total number of observations is a type of success rate. This enables us to conduct a standard binomial test to compare the success rates. Specifically, we can test the accuracies of our models against some accuracy of a benchmark classifier and we can further understand whether our models can provide more information than what the benchmark classifier provides.

### 2.2.4 Selection of Benchmark Classifiers: On the No Information Benchmark and Martingale Benchmark

Since the efficient market hypothesis suggests that the market is efficient and there does not exist a relation between future returns and historical information, price changes are decided with new information to be released. Consequently, we want a benchmark classifier that makes predictions based on limited information or no information to reflect the implication of the efficient market hypothesis. In general, there are two types of classifiers we can consider for a no information benchmark. First, we can consider a random classifier. Second, because the investors are rational and the prices in the efficient market reflect the expected future prices, investors have a rational understanding of the distribution of prices as required by the consistent beliefs which lead to market efficiency. Thus, we can consider a classifier with distributional information of returns. Discussion about the inclusion of classifiers with distributional information of returns is provided in the next subsection.

We want to be conservative. Therefore, from those classifiers using limited or no information, we want to select a classifier that produces the highest possible accuracy. We consider 5 candidate benchmark classifiers. First, we consider a random classifier that takes into account absolutely no information. The random classifier labels each OOS observation with a random label from return state 1 to return state 10 with equal probabilities. The accuracy level of the random classifier reflects the situation where no relation exists between future return states and historical information. Second, we consider a random classifier that randomly labels OOS observations with probabilities based on the observed IS return state probability mass function. Third, we consider a naive classifier that labels OOS observations with the most populated IS return state. The second and the third benchmarks represent the situation where the market knows the IS distributional information of return states. The machine learning literature argues that if the distributional information of the response variable is the only information or the only useful information, classifying all observations predictively into the majority class is the best guess. Our multiple comparison tests confirm this argument. Fourth, we consider a random classifier that randomly labels OOS observations with the knowledge about OOS return state distribution. Fifth, we consider an accuracy by a naive classifier that labels OOS observations with the most populated OOS return state. The fourth and fifth benchmarks are enhanced versions of their IS counterparts.

Beyond the 5 possible classifiers to produce the no information accuracy benchmark, we also consider a martingale accuracy classifier under the assumption that the returns follow a memoryless process. Under this assumption, the best prediction for the future return state is the current return state. Therefore, we specify a classifier using the current return state as the prediction for the future return state. The accuracy produced with this martingale classifier is our martingale accuracy benchmark.

### 2.2.5 Discussion on Classifier with Distributional Information of Returns

There are three reasons for us to consider the benchmarks reflecting the distributional information about return states. First, it is a convention in machine learning literature to test predictability against the naive classifier as a benchmark. Since the distributional based classifier accuracy is easily accessible through our empirical observations and does not involve any modeling or predictors, it is a natural choice to examine whether a model and the associated predictors are working, i.e., whether the model and the predictors provide more information than the distributional information. We follow the convention, considering that the distributional information is the minimum amount of information. The accuracy delivered by a naive classifier is often referred to as no information accuracy indicating that the minimum amount of information is reflected by the accuracy. Therefore, the benchmarks including the distributional information do not deviate us from our testing purpose about whether useful information is provided by our models and the historical information that our models use.

Second, we would like to demonstrate how we select the most restrictive benchmark among the candidate classifiers for our binomial tests. The selection is conducted through a multiple comparison test with the Monte Carlo samples of accuracies by the random classifiers and the naive classifiers. The test looks into which classifier provides the highest accuracy on average. The OOS prediction accuracies by each of the random classifiers and the naive classifiers provide the simplest setup to test predictability within our novel testing framework. The selection process contributes unique insights to the market efficiency literature. In fact, through Monte Carlo simulation and the multiple comparison tests, we show that the naive classifier using the IS distributional information about return states delivers an accuracy that is statistically higher than what is delivered by the random classifier that uses absolutely no information. This means that the future returns can be better predicted by the random classifier using the basic information about IS probability mass function of the return states. However, it is worth noting that predictability does not necessarily equal the rejection of market efficiency. A meaningful empirical question about market efficiency from predictability has to come with meaningful profitability. We do not see such profitability with the naive classifiers. We will discuss more this point in the subsection below.

Specifically, for each iteration of the Monte Carlo simulation, the random classifiers and the naive classifiers predict the testing data set return states according to the mechanism mentioned above. In total, in each iteration, the Monte Carlo process samples 4886 return states from the real data set mimicking the average number of stocks in each month of our entire sample. We iterate the simulation 10,000 times. The accuracies for each classifier then allow us to conduct multiple comparison tests. We introduce Tukey's HSD and demonstrate the Tukey's HSD test conducted with a time period covering 199201:201912 as the testing sample in Table 3. Tukey's HSD confirms that the naive classifier with OOS distribution knowledge provides the highest accuracy on average. We select the naive classifier as the benchmark to further test whether our models indeed provide information about the relation between future returns and historical information in the binomial tests.

Third and most importantly, a classifier that considers only the distribution of the return states mimics the investor behavior under the assumption of the market efficiency and the associated accuracy is the benchmark as implied by the theory. An efficient market means that the future returns are unpredictable based on historical information and that investors are rational with enough information about the distribution of returns. To reflect the investor behavior, the inclusion of the benchmark based on a naive classifier that considers the distribution is necessary and meaningful. In fact, according to the rational investor assumption which leads to the efficient market hypothesis, investors have consistent beliefs (Sargent 1993, Barberis and Thaler 2003). Consistent beliefs mean that investors have correct information about the distribution they use to forecast unknown state variables. In other words, investors in an efficient market must have enough information to infer the distributions of state variables, including the distribution of returns. Therefore, more precisely, we will want to include the accuracy based on the true OOS return state distribution as it is the best distributional information the rational investors can infer. We generally refer to the accuracy provided by naive classifiers with distributional information as no information accuracy hereafter. Strikingly,Table 3 show, because of introducing historical return state information, Classifier 6 has better overall accuracy in our simulation. In the binomial test, we thus include both Classifiers 5 and 6 as our benchmarks in our binomial tests.

### 2.2.6 Binomial Test: A Joint Test

The binomial test is a joint test and provides unique economic insights to finance machine learning literature and market efficiency literature. Specifically, given any one of our models built based on historical public information to predict future return states, if the prediction accuracy is tested as significantly greater than the no information accuracy, the statistical significance suggests at least 3 aspects.

1. We can conclude that the prediction accuracy of the model is statistically better than the no information accuracy by the naive classifier. In other words, the combination of the model and the predictors delivers good predictive performance.

2. Since the naive classifier using only basic distributional information provides minimum or no information, the statistical significance of better accuracy indicates that the associated model delivers statistically meaningful information and this information is beyond the basic distributional information. Furthermore, because of using historical information and the specific modeling structure, the meaningful accuracy signals that there is a relationship that exists between future return states and lagged predictors and that the relationship is at least partially decoded by the related model.

3. If a binomial test presents significance and the associated portfolio strategy can generate profits, then the prices may not be correct. The significant predictability suggests that there exists a relation between the predictability and the information uncovered by the model using historical information. The predictability proves that the future prices will move towards the predicted level and the profitability suggests that the piece of information uncovered by the model is useful. In other words, the historical information can generate future profits. As mentioned by Fama (1991), the return predictability does not necessarily mean that the market is inefficient. If the predictability does not allow the generation of profit, the market is efficient in terms of allocating resources and the prices correctly reflect information. Therefore, the predictability has to be combined with the meaningful profitability that cannot be offset by friction. We show that our models generate large economic gains for investors (from the previous section) and the predictability is significantly based on our modeling architectures and historical information.

An important implication of profitability based on historical information is that the market by large does not share the information. If most of the market participants share this information generated by our models with historical information, the arbitrage process will erase the opportunity to benefit from the model predictions. Therefore, the information generated by our models provides new information about future returns and historical information. At the same time, at least to our models, the market price may not be correct historically for the CRSP-COMPUSTAT sample covering 196301:201912, since correct prices should not allow any investor to generate new information using historical information and apply the generated information to make OOS profit. In summary, the binomial test is a joint test of statistical significance of predictability and market efficiency. We discuss more the use of the binomial test in Section 3.

## 2.3 Data Construction and Sample Splitting

### 2.3.1 Data Components

Our data universe contains 3,342,486 monthly stock observations of 26,302 distinct common stocks with current returns listed on 3 major exchanges covering the time period of 196301:201912. We present the basic summary statistics of our data in Table 4. We construct a modeling sample of 332 lagged predictors. The lagged predictors include the return state, 101 firm characteristics, 2-digit SIC industry indicator, 2-digit SIC industry lagged returns, 125 macro indicators. We augment the macro indicators with 9 market-specific predictors based on Goyal and Welch's data set. We also sort the past 94 numeric firm characteristics and further augment the macro indicators with differences between the top decile median returns and the bottom decile median returns.

Specifically, we fully reconstruct the firm characteristics based on Green et al. (2014) with CRSP and COMPUSTAT. We make the data set to be completely CRSP centric data with no data elimination if possible. In the end, we only eliminate rows with missing current returns and the securities that are not common stocks with SHRCD of 10, 11 or 12 listed on the major 3 exchanges with EXCHCD of 1, 2 or 3. Figure 1 presents our sample coverage of the CRSP universe with the counterpart reference level based on the entire CRSP database. Note that the reference level includes securities with missing returns and securities that are not common stocks.

Following Green et al. (2014) and Gu et al. (2020), we lag the annual firm characteristics by at least 6 months, we lag the quarterly firm characteristics by at least 4 months and we lag the monthly firm characteristics by at least 1 month. Note that the firm characteristics include variables depending only on historical returns, such as return momentum, and the variables depending on financial information and corporate announcements such as earnings and IPO. This data set of firm characteristics provides a sound experimental environment for the test of market efficiency using machine learning methods. The lagged characteristics ensure that all the information in our models is historically publicly available and there is no forward-looking information leak in terms of the return based information nor the corporate announcements.

We also include macro time series indicators to allow the models to learn about the relative intertemporal position of the market. We obtain the macro indicators of McCracken's fairly new FRED-MD database from the website of the Federal Reserve Bank of St. Louis. The database provides most of the mainstream macroeconomic indicators and is updated on a monthly basis (McCracken and Ng 2016). We retain 125 of the macro indicators in the end and exclude the variables ACOGNO, ANDENOx, and TWEXMMTH due to limited availability. We obtain Goyal and Welch's (2008) data set from Amit Goyal's website. We include d/p, d/e, svar, b/m, nits, corpr, dfy, dfr and ltr in our final data. We lag the time series macro indicators by at least 1 month.

For factor model tests, we obtain Fama and French's (1992) MKT-RF, SMB, and HML factors augmented with MOM factor of Carhart (1997). We collect these factors from Kenneth French's data library. We also include the Hou, Xue and Zhang's (2015) q 4 factors — R_ME, R_IA, R_ROE as well as the R_EG factor from the update of Hou, Mo, Xue and Zhang (2018). These factors are collected from Lu Zhang's investment CAPM website. We include these pricing factors in the factor model tests mainly because of their good performance in the empirical asset pricing literature and their popularity.

We execute our models for both of the data setups and notice that the OOS results especially for ANN models are not very different with or without the macroeconomic data components. Interestingly, the performance for tree models increases with less macro-level predictors suggesting the contribution for individual stock return state prediction is mainly from characteristics. Therefore, despite that we have included all the popular data components for individual stock return studies, most of our empirical results are the results produced with the data set excluding the macroeconomic data components, i.e., Goyal and Welch predictors, 94 single sort factor mimicking portfolios and FRED-MD macroeconomic indicators. Table 1A demonstrates the comparison.

### 2.3.2 Data Manipulation and Sample Splitting for Training and Testing

We manipulate the data to obtain the most appropriate modeling inputs according to conventional machine learning model requirements. First, scaling variables is an important step to minimize the impact on weight adjustment and tree split because of the scale difference. In the machine learning literature, scaling variables is almost always recommended. We follow this standard practice. We scale the numeric firm characteristics cross-sectionally through normalization in each trading month if the field is not missing. We scale firm characteristics cross-sectionally for each time point to capture the relative characteristics difference at a time point across stocks.

Second, we also follow the conventional practice in machine learning literature and fill the missing values in categorical firm characteristics by specifying a new category. This practice does not bias the data set. Instead, it keeps the observations and helps the model to split more indicator specific effects from the general effects by retaining more observations. Third, after scaling, we fill the numeric firm characteristics with 0, which is the cross-sectional mean and median in each trading month. We also fill the time series variables, including the macro indicator variable consumer sentiment index (UMCSENTx), with the latest available values. Despite that there is the possibility of introducing noise, we fill the missing values because the possibility of including more useful information due to retaining more observations is also high. In the modeling phase, we scale the time series predictors through normalization with all available IS history such that the time series predictors can capture the intertemporal relative position of the market trend and the macroeconomic trend.

To enlarge the testing power, our data covers the time window of 196301:201912, almost everything in the CRSP-COMPUSTAT universe. We split the entire data set in two ways corresponding to our time series tests and the purpose of our cross-sectional tests. For our time series tests, specifically, we want to test on intertemporal applicability of the models fitted with our time series in-sample training process in the OOS periods. Therefore, we split the data in the middle based on the time length. As we have 57 years of data, we choose the end of the year 1991 as the splitting point as the earlier data has less number of stocks on average. Then, we have 2 time series subsamples, the first subsample covers 196301:199112 and the second subsample covers 199201:201912. Later, we refer to the subsamples in terms of the coverage of periods.

To take the advantage of the long time window, we train our models with the subsample covering 196301:199112 and make predictions with the subsample covering 199201:201912. We also train our models with the subsample covering 199201:201912 and make predictions with the subsample covering 196301:199112. We present the results of the OOS predictions with economic metrics both in subsamples and combined. We can use the latter 28 years to make predictions on the first 29 years, because we focus on extracting the relation between the historical public information and the future return states and our models do not depend on time structure. Therefore, we will not violate the setup of OOS tests.

It is obvious that the current working rules apply to the next period while it is not obvious whether the current working rules can apply to the past. Training models in latter dates and testing model in former dates makes a strong case of testing the time-invariant applicability of the models. For example, after the appearance of the Black Scholes and Merton (BSM) model, we know that BSM will certainly work for the pricing of stock options in the future after its appearance. On the other hand, it is hard to tell whether BSM would work to the dates prior to its existence. If a model captures the real process of return state transition in general, the model should be applicable to any time period.

Our time series sample splitting focuses on the evaluation of time series OOS tests, while we position our cross-sectional sample to focus on overall explanatory ability across the entire time series. Unlike traditional econometrics, it is not ideal to test a machine learning model with the sample that the model is trained with. To overcome this issue, we split the data by odd number months and even number months in the spirit of Fama and French (2018). Our cross-sectional data split permits us to model the entire length of the data set across different macroeconomic conditions and test the fitted models in an OOS testing setup with completely new observations that are not seen by the models. We summarize the sample splitting for training and testing in Table 4.

# 3 Time Series OOS Performance

In this section, we present the OOS performance of our classification-based portfolios. We show that our portfolios perform systematically better than what the market portfolio based on buy-and-hold strategy can deliver, regardless of the weighting schemes and the market capitalization cutoff points. We compare the performance in the aspects of average return, return volatility, risk-return tradeoff, and shortfall. Our demonstration of the superior OOS performance of the classification-based portfolios answers Thaler's first question about whether we can beat the market. We further demonstrate the factor model explanation of the returns generated by the classification portfolios and show that the traditional factor models cannot explain the returns generated by our portfolios.

To carry out the OOS economic metrics, we first train our time series models using the first half of our sample covering 196301:199112 and make predictive classifications on return states in the second half of the sample covering 199201:201912. We then train on the second half of the sample covering 199201:201912 and make predictive classifications on return states in the first half of the sample covering 196301:199112.

As discussed in the sample construction section, we can use the latter half of the sample to predict the former half of the sample without any violation of the OOS test setup. By always using lagged information to score models, the models trained with the time series subsample covering 199201:201912 do not see any OOS observations, if we test the models with the time series subsample covering 196301:199112. We combine the OOS predictions in the two windows to form portfolios covering the entire sample period of CRSP-COMPUSTAT universe and provide by far the longest time series OOS test coverage.

As courtesy insurance of the model robustness, we provide performance tables of separated periods in the appendix[5] along with the results from the portfolios enforcing strict market capitalization cutoffs. Neither splitting the full length OOS metrics into two halves nor enforcing the market capitalization cutoffs changes our conclusion about the superior OOS performance of our classification portfolios comparing to holding the market.

After the demonstration of the good time series OOS performance that is time-invariant, we examine the statistical metrics of the accuracy that drives the good economic performance of our portfolios. We show that our portfolios can deliver balanced accuracy that is good comparing to the referencing levels of the class prevalence in our data. We further test the overall accuracy of our models against the no information accuracy of the naive classifier through the binomial test. The test against the no information accuracy as the null hypothesis is a conventional test in the machine learning literature. Following the important interpretation of the binomial test discussed in Section 2, we show that the OOS overall accuracy levels provided by the predictions from our models are statistically significantly larger than the no information accuracy, which indicates the generation of new information about OOS return states based on predictions from the models fitted in the training process. We discuss the importance of this test to the understanding of market efficiency. We further look into the prediction accuracy by return state transitions and illustrate the opportunities captured by our models from learning with IS historical observations. In the end, we discuss the by-class statistical metrics, the training performance, and the model selection during training.

### 3.1 Performance Evaluation with Economic Metrics

Because of the large number of models under consideration, we rely on visualization to summarize the major results and provide the performance metrics in table format in the Appendix. Figures 2 and 3 present the summarized economic metrics of the classification portfolios covering 196301:201912. The dashed blue line is the performance provided by the market portfolio, i.e. buy-hold strategy applied to the entire market. The long portfolio indicates a long position in all stocks predicted to be in the return state 10, or the best return state. The short portfolio indicates a long position in all stocks predicted to be in the return state 1, or the worst return state. The long-short portfolio includes a long position in the return of all stocks predicted in the return state 10 and a short position in all stocks predicted in the return state 1.

Figure 2 presents the OOS performance of the equal-weight portfolios and Figure 3 presents the

[5] We have the results but have not included them in the appendix yet. They will be included soon.

OOS performance of the value-weight portfolios. Note that we change the weights in the long leg and the short leg separately when implementing a long-short portfolio. The allocation to the long leg and the short leg in a long-short portfolio is always equal, i.e., having a 100%:100% allocation ratio, across weighting schemes. All the returns are fully risk-adjusted against the risk-free rate and the cumulative returns are calculated as the gross returns net of the initial investment.

The first row of Figure 2 (Figure 3) demonstrates the distributional information of the equal-weight (value-weight) portfolio returns. The long portfolios deliver better monthly average returns comparing to the market monthly average return. The short portfolios deliver worse performance by taking the long position in the predicted worst-performing stocks. When combined, the long-short portfolios deliver systematically better average monthly returns comparing to the market return.

Over the full-length OOS performance evaluation period covering 196301:201912, our neuron networks and tree models successfully distinguish the best-performing stocks and the worst-performing stocks. A similar conclusion holds in the comparison of the standard deviation. Our long-short portfolios deliver surprisingly small monthly return standard deviations. In the comparison in the return skewness against the market, our portfolios seem to push the skewness towards positive values. This implies that our portfolios are more likely to realize large gains than large losses. However, if one is interested in short only strategies, this means more significant loss. The kurtosis plot demonstrates that our portfolios deliver returns that are with a more concentrated center, which indicates a reduction of the distributional fat tails. We augment the first row in Figures 2 and 3 with Figures 4 and 5 presenting return distributions for equal-weight portfolios and value-weight portfolios respectively. Since the return distributions of the long portfolios are to the right of the return distributions of the short portfolios, Figures 4 and 5 further confirm that our models can distinguish the future best return state stocks from the future worst return state stocks. It is also worth noting that the return distributions are closer to normal for our classification based long and short portfolios.Specifically, Figure 4 shows the best OOS equal weight long-short portfolio monthly average return in our test is delivered by our neural network model ANN2 32. The long-short portfolio delivers an average of 3.4% monthly return from 196301:201912. The long-short portfolio based on our tree model, GBM8 100, delivers an average monthly return of 2.2%. During the same period, the market portfolio delivers an average monthly return of 1.1%. Figure 5 conveys similar message. For the value weight scheme, the long-short portfolio based on our neural network model ANN2 32 delivers an OOS average monthly return of 1.92% from 196301:201912, while the market portfolio delivers an average monthly return of 0.9%.

In Figures 2 and 3, the second row demonstrates SR and CEQ of the full-length OOS evaluation covering 196301:201912. Because of holding the predicted worst-performing stocks, short portfolios deliver negative SRs. The long portfolios based on complex models deliver SRs that are higher than what the market provides. Our long-short portfolios deliver surprisingly high SRs achieving systematically superior performance in terms of the risk-return tradeoff. The best performing long-short portfolio based on our gradient boosting machine with the depth of 8 delivers an SR of 0.81, when we implement equal weight portfolio formation. The SR drops to 0.42, when we implement value weight portfolio formation, still more than doubled the SR delivered by the market. The drop in performance is understandable as we cannot gain the returns from the small-cap stocks with a 1:1 ratio. However, this does not mean that the portfolio performance is completely driven by the microcap stocks. In the results not reported, we check the performance with strict capitalization cutoffs across all metrics, the performance of the sample eliminating the bottom 5 % capitalization stocks and the performance of the sample eliminating the bottom 10 % capitalization stocks are both comparable to the results covering the entire sample, especially for value weight portfolios. We also have an extremely conservative set up to show the robustness in Section 4. Note that our SRs are calculated with monthly returns instead of annual returns and are not adjusted with modeling R squares, while Chen et al. (2020) use annualized returns in the calculation of SRs and Gu et al. (2020) use modeling R squares to adjust their SRs. Annualizing the returns can smooth the time series and shrink the size of the variance while adjusting the SR with R squares can magnify the size of the numerator and shrink the size of the denominator. Both of these modifications can inflate the SR. To annualize our SRs, we need to multiply the SRs with $\sqrt{12}$.

We also present the CEQ and overall cumulative returns (CAR) in natural logs during the full-length OOS evaluation period. In terms of economic investor gains, both the CEQs and the cumulative returns show the superior overall performance of our long portfolios and long-short portfolios compared to what the market can generate during the 57-year investment window. The reason for the similar look of the CEQ subgraph and the cumulative return subgraph is that the construction of the log scale cumulative return is similar to the construction of CEQ. In the evaluation of equal weight portfolios, our best performing long-short portfolio based on our ANN model of 2 hidden layers and a total of 32 neurons deliver a full-length period cumulative return of 485,649,224,179 % net of the initial cost. In the evaluation of value weight portfolios, the cumulative return of the long-short strategy drops to 20,738,716 %, which is still 10,000 times what the market achieves with value weights. We augment the cumulative return evaluation in Figures 2 and 3 with Figures 6 and 7. As demonstrated, the cumulative returns of our long-short portfolios surge to a level that both the market cumulative returns and the cumulative returns from holding the predicted worst-performing stocks become flat lines.

Figure 6 (7) demonstrae our equal-weighted (value-weighted) long-short portfolios deliver phenomenal cumulative returns in the OOS test. Comparing to our long-short portfolios, the market portfolios is a horizontal line in the subgraphs as the cumulative return delivered by the market over the same period is too small. Our Neural Network model, ANN2 32 OOS delivers a cumulative return of 485,649,224,178.58% (20,738,715.94%,). The long leg of ANN2 32 alone delivers a cumulative return of 29,138,823,156.42% (xxxx) from 196301:201912. The equal weight long-short portfolios based on our tree models, GBM8 100 and GBM6 100, deliver OOS cumulative returns of 200,202,148.41% (2,653,376.44%,) and 64,231,407.4% (1,988,854.31%) respectively during the same investment period. During the same period, the market portfolio achieves a cumulative return of 72,874% (27,396%), substantially lower than what is delivered by our tree model DRF4 200 (DART2

100). Despite that DRF4 200 (DART2 100) is clearly under-fitted, it still delivers a cumulative return of 84,804.67% (429,354.55%).

After the surprisingly good performance demonstrated in the return distributions, the risk-return trade-offs, and the economic gains, we look at the largest shortfalls associated with our classification portfolios. We evaluate the shortfalls with the maximum drawdown. The short portfolios are associated with the smallest maximum drawdown. Our long-short portfolios deliver a surprisingly low level of maximum drawdown. In the equal-weight implementation, our best performing long-short portfolio in terms of maximum drawdown is again the long-short portfolio based on the gradient boosting machine GBM8 100. It achieves a maximum drawdown of 11.55 %, only 15 % of the market maximum drawdown, which documents the historic selloff around the financial crisis. When switching to value weights, the maximum drawdown is still meaningfully lower for the long-short portfolio based on GBM8 100.

## 3.2 Explanatory Analysis with Factor Models

We further provide alpha as an additional metric of the OOS performance evaluation and look into whether the performance can be driven by common market factors. We regress the return time series of our classification portfolios on the popular factors. We include Fama French 3 Factor model (FF3F), Fama Frech 3 Factor + MOM model, the original q-factor model (q4) of Hou, Xue and Zhang (2015) and the q5 factor model of Hou, Mo, Xue and Zhang (2018) with the R_EG factor.

Figure 8 and Figure 9 summarize the factor model tests with our equal weight and value weight portfolios respectively. The orange dashed line indicates the position of zero. The first row shows the alphas, the second row shows the t values associated with the alphas and the last row shows the R square values associated with the factor models. In the case of equal-weight portfolios, clearly, FF3F models and FF3F + MOM models cannot explain any of the long-short portfolios based on our classification models, not even the long-short portfolios generated by our underfitted tree models with a maximum depth of 2. Despite that the q factor models perform slightly better, the q factor models can barely explain only the returns generated by the long-short portfolio based on the clearly underfitted DART2 100. The conclusion with the long-only portfolios is similar. In fact, the alphas of our portfolio returns are associated with the t values as large as 20 in the equal-weight implementation of the long-short portfolios. The R squares in many cases are as low as 1 % for our long-short portfolios, indicating the associated factor models do not explain the variation. The factor models' ability to explain the long-short portfolio returns does not get any better systematically when we shift from equal weight to value weight.The factor models cannot explain the returns achieved by our long-short portfolios. Taking our long-short portfolio based on our neural network model, ANN2 32, the portfolio has the alphas of value -weighted (equal-weighted) portfolios as of 3.31% (1.9%), 3.37% (1.85%), 3.35% (1.84%), 3.26% (1.71%) respectively against FF3F, FF3F + MOM, q4 and q5. The explanatory power of the factor models on the portfolios also seems decreasing as the complexity of our classification models increases.

## 3.3 Model Complexity and The Economic Performance

Gu et al. (2020) mention that in their implementation targeting the numeric values, they find the shallow learning outperform deeper learning. They document that the performance of the neuron networks peaks at three hidden layers and their tree models tend to select trees with few leaves. Our performance evaluation with the economic metrics shows different conclusions.

First, the performance of the neuron network is largely influenced by the total capacity of neurons instead of just the number of layers. Controlling the number of layers, adding neurons improves the performance. In the factor tests, we do not see a significant deterioration of performance when we increase the number of layers, either. Second, for our tree models, there is an obvious improvement trend in all of our subgraphs in Figures 2, 3, 8 and 9. As we increase the maximum depth of our tree models, the performance of our tree model-based portfolios improves substantially. Specifically, with a maximum depth of 2, our tree models seem significantly under-fitted and the associated long-short portfolios can deliver performance metrics that are worse than the counterparts of the market, while our tree models with a depth of 8 can form the best performing portfolios. Third, for each of our models, as we increase the model complexity, we can see a widened gap between the long portfolios and the short portfolios across figures. This shows that the models can better distinguish the future best return state stocks from the future worst return state stocks as we increase the model complexity.

## 3.4 Performance Evaluation with Statistical Metrics

We have demonstrated that the classification portfolios can beat the market and the superior performance of the classification portfolios in a range of economic metrics. In this subsection, we analyze the classification models behind the good economic performance through statistical metrics. We first show the OOS classification accuracy achieved with our models and test the OOS classification accuracy against the true OOS no information accuracy delivered by the naive classifier. The no information accuracy by the naive classifier is under the assumption that the response variable, the future return state, is independent of the lagged predictors. We also dig into the by-class level accuracy, analyze the overall return transitions and provide new insights about market efficiency. We show that our machine learning models take the advantage of the difference in the information uncertainty to deliver superior statistical performance. We base our discussion in this subsection on the combined OOS predictions covering the entire time series. We provide the tables for separate periods in the appendix for robustness check.

### 3.4.1 Accuracy and Binomial Tests Against the No Information Accuracy

The accuracy metric is a unique and natural statistical metric that is available to the classification problems. In numeric predictions, one can only construct metrics based on the prediction errors as it is impossible to measure accuracy directly. The error-based metrics are hard to comprehend. However, in the classification problems, all of the predictions are similar to combining problems as the predictions are by nature similar to choices and selections. Using accuracy calculated as the percentage of correctly predicted return states, we can directly measure the prediction quality.

Column 2 in Table 7 presents the OOS accuracies, Accuracy, of our models that predictively decide which stock is more likely to be among the best-performing stocks and which stock is more likely to be among the worst-performing stocks. We also provide Kappa statistics associated with each of the models in Column 2. The models delivering good economic performance are systematically better performing in terms of the OOS prediction accuracies as well. For example, the underfitted DRF2 200 model delivers an accuracy level of 15.2 %, while the GBM8 100 model, whose associated portfolios are among the best performing portfolios, delivers an accuracy level of 15.8 %.

Beyond the direct interpretation of the statistical performance, we adopt the classification framework to take the advantage of the accuracy metric as a possible direct proxy that can be used to conduct a statistical test on whether we can foresee the future with historical information. We further introduce the binomial test to the efficient market hypothesis literature and the financial machine learning literature. As mentioned in Section 2, since the accuracy metric is in a natural form of proportion, we can compare the statistical difference between 2 accuracies through a binomial test. Specifically, we test the predictive classification accuracies against the accuracy delivered by the naive classifier as the null hypothesis.

The binomial test on accuracy has profound meaning in our setup.

a) iit tests the statistical meaningfulness of the accuracy delivered by a classification model. If a model delivers an OOS accuracy that is statistically significantly higher than the no information accuracy, the model captures the statistically meaningful essence of the relation between the future return states and the historical information. In other words, there exist useful contents in the historical information about the future return state and our models are able to extract the useful contents. We show in Table 7 that all of our models deliver statistically significant OOS accuracies in tests against the no information accuracy.

b) To the finance literature, the accuracy is a natural statistical metric to evaluate the market efficiency and has special meaning to us as discussed in Section 2.

Economically, if the market is efficient in the strong form, then the prices fully reflect all available information, regardless of whether the information is private or public. In this situation, we should not be able to benefit from trading stocks using any historical information.

If the market is efficient in the semi-strong form, then the prices fully reflect all historical public information. An investor can only benefit from trading stocks using private information. In other words, even if we can generate profits in the semi-strong form efficient market, the historical public information should not contribute to the prediction of the returns and only the private information can contribute to the prediction of the returns.

If the market is in the weak form of efficiency, we can benefit from trading on the information included as the public information, including the corporate announcements and the historical macroeconomic information, will be incorporated into the prices gradually.

However, if the market is efficient in the weak form, past returns should not contribute to the prediction of future returns. If the market is not efficient at all, the returns are predictable and all types of information can contribute to the prediction of returns.

In summary, the binomial tests on the prediction accuracies, coupled with the analysis of the predictor contribution, can provide direct indications about whether there exists a violation of the 3 forms of market efficiency. Statistically, we should not observe any meaningful OOS accuracy in our predictions, if the market is efficient in the strong form. However, if the prediction accuracy of a model is statistically significantly higher than the no information accuracy, we can conclude that there exists some relation between the historical information and the future return states, and the relation is captured by the machine learning model. In other words, the statistical significance of the accuracy can bring up at least the questions about the market efficiency in the strong form. Therefore, what we present with the binomial test is a direct test of the strong form market efficiency. The contribution of historical predictor variables in the predictability can further help us understand whether we should further question the semi-strong form and the weak form of market efficiency.

We present the binomial test results against the OOS no information rate with the last 4 columns in Table 7. As mentioned in Section 2, the no information accuracy measures the best prediction one can make if the predictors do not have any relation with the response variable in the classification problem. We make it stricter by using the real ground truth OOS no information accuracy based on the OOS return state distribution.

The binomial test through accuracy against no information accuracy is a test about whether the response variable and the predictors are independently distributed. Therefore, the no information accuracy provides a natural tool for us to check if our models are providing OOS information predictively based on the historical information including return information, corporate announcements, and macroeconomic indicators. In an ideal case, we want to split our returns cross-sectionally by trading month into evenly distributed classes. However, because we have the thresholds of the quantiles when splitting our returns, each return state includes a slightly different number of stocks cross-sectionally within each date and the difference remains there when we look at the entire time coverage. 10.16% is the portion that the return state 7 takes in our entire sample across 196301:201912 and the return state 7 makes the largest portion of our sample among all return states. Therefore, after observing the entire distribution of the return states, an investor should always bet on return state 7, if the return state transition is truly independent of any of the historical information associated with EMH.

It is obvious that the prediction accuracies delivered with the models trained IS based on historical information successfully surpass the ground truth no information accuracy. With the provided 99% confidence intervals and the p values, it is clear that the accuracies delivered by our models are statistically significantly higher than what the ground truth return distribution can deliver. Without looking into the contribution of predictors, which will be discussed in Section 5, our findings in Table

7 has two important implications. First, statistically, we confirm that there is some relation between the OOS return states and the IS lagged predictors, including the historical return information, the historical corporate announcements, and the historical macroeconomic indicators. In other words, as the response of the model, the future return states are not independent of the historical information. Second, economically, we confirm that the market efficiency can be improved as the return states are predictable at a significant enough level through the existence of the relation between the OOS future return states and the historical information. The predictability shows that the historical information is not fully reflected by the prices.

Based on the fact that our models deliver significant OOS accuracies, assuming that between two investors, one investor uses our models and the other investor believes that the return states are independent of historical information and thus relies on the OOS true distribution information of return states to make predictions. Our finding means that the investor using our model can then make OOS return state predictions that are better than the best guess that the other investor can make after she sees the entire ground truth OOS distribution of the return states.

Our model captures the relation between historical information and the OOS future return states, and thus generates new information that the prices do not reflect. In other words, the investor using our models have an information advantage and manually introduces de facto information asymmetry to the market. In practice, this implies that sophisticated investors, such as Renaissance Technologies, can take the information advantage by generating new information predictively about the future return states based on historical information. In general, the generated information, depending on the choice of analytical tools and the sophistication levels of the investor, may not be publicly available even if the other investors can observe the ground truth distribution of the future return states. We discuss more the semi-strong form market efficiency and the weak-form market efficiency in Section 5 based on the predictor contributions.

### 3.4.2 Training and Model Selection

With the performance in OOS economic metrics and statistical metrics, we can learn about the applicability of the models in real practice by looking into the training process and see if we can select the correct models to generate the guideline for portfolio allocation. We briefly discuss the performance metrics of the training process in this subsection.

Figure 10 and Figure 11 show the economic performance of equal-weight portfolios for the 2 time series training sets covering 196301:199112 and 199201:201912 respectively. The IS economic performance of our classification portfolios are similar to their positions in the OOS economic performance evaluation. Our best performing portfolios are still the models with more complex structures, such as GBM8 100. Table 16 summarizes the statistical metrics of the IS performance of our models for the 2 time series training sets. The accuracies of our models are substantially higher than what we have for the OOS evaluation, which is expected. The best performing models are also the models with more complex structures. The GBM8 100 can deliver accuracy levels around 21% in the two training sets respectively. The good performance and the consistency between our IS and OOS models ensure the applicability of our models in real practice. In other words, through the training evaluation, we can rank the performance of our models correctly and choose the selected model to make OOS predictions.

# 4 Behind OOS Performance: An Inquiry into the Economic Insights

## 4.1 Return State Transition Uncertainty and the Prediction Strategy of Machine Learning Models

We next assess the by-class statistical performance of the OOS predictions. We first present the ground truth about return state transitions and discuss the relative uncertainty across the return state transitions. We then evaluate the prediction accuracy level for each individual return state transition achieved by our classification models in general. We demonstrate the strategy that our models take to achieve superior OOS performance and discuss the easiness of predictions. We further present the by-class OOS performance for each model with popular classification related statistical metrics at the end of the section.

### 4.1.1 Return State Transition Uncertainty, By-Transition Performance and Implication on Market Efficiency

Table 8 presents the information of the ground truth return state transitions. From Panel A, we can see that the return state transition probability based on the entire sample covering 196301:201912 is not evenly distributed. The extreme return states and the middle return states are associated with transition probabilities either substantially greater than 10% or substantially lower than 10%. These states are thus with higher certainty in the process of state transition. More specifically, return state 3, return state 4 and return state 9 seem the most uncertain states. The return state 1 and return state 10 seem the most certain states. In general, the corner transitions in the matrix, such as the transition from the current worst performing state to the future highest performing state, are associated with substantially higher probabilities. The results seem to support short-term momentum (state 1 staying at state 1 with 17% probability and state 10 staying at 10 with 12.3% probability) and short-term reversal (from state 1 to state 10 with 18% probability and state 10 to 1 with 17%) strategies. While the center transitions of the matrix are associated with more evenly distributed probability around 10%. It is also clear that the transitions from the current middle level performing return states to the future extreme performing states are with relatively lower probabilities.

Panel B presents the monthly mean returns of the associated return state transitions of all stocks in our sample covering 196301:201912. We can see that the better future performing states are with better cross-sectional average returns. However, the panel shows that the transitions from the current middle-level return states, i.e. return state 2 to return state 9, do not really deliver very different average returns when they transit from the current return state to the new return state. For example, the average return for a stock transiting from return state 3 to return state 9 is not very different from the average return of a stock transiting from return state 4 to return state 9. However, the corners of the mean return table show a different situation. A stock transiting from return state 1 to return state 10 delivers a significantly higher average return compared to a stock transiting from return state 2 to return state 10. Both Panel A and Panel B show that the extreme return transitions are where the opportunity lays for the investors who want to generate excess trading profit. We show that this is exactly the strategy of our models.

Table 9 presents the average accuracy of time series OOS prediction across our models. Our models put a high stake in the extreme return state transitions. The transitions to the lowest return state obtain the highest overall accuracy levels in OOS prediction. The models also pay more attention to the transitions to the middle return states and the best performing return states. The OOS predictions from our models seem successfully capturing the better certainty of the transitions to the extreme states and the transitions to the middle performing states. This implies that our classification models believe that the extreme return states and the transitions to the middle return states are more predictable. This is consistent with our observation on the true transition probability matrix.There are about the same probability and accuracy that stock in state 1 will move to state 10 and remain in state 1. With 5% lower probability, a strategy of holding stocks in state 10 is also attractive; however, the accuracy of this strategy is much lower than the other strategies. Our model consistently suggests a contrarian strategy of shorting stocks that are in state 10 is the best strategy. It generates the high probability and accuracy with returns of 23.21%.

As our models capture this difference in the levels of uncertainty during IS training process and are able to tell what return state is more predictable, the OOS prediction accuracy by return state transition shows the models' inclination about what stocks are priced more efficiently under the current market condition. Our models gain from the inefficiently priced stocks, because those stocks deliver more certainty about return state transition through unevenly distributed transition probabilities. Based on Table 9, our models clearly have a systematic preference of betting on the extreme transitions and the transitions to the middle return states. Therefore, specifically, the accuracy by return state transition, combined with the ground truth difference in the probability of transition, indicates that the extremely priced stocks and the middle performing stocks are creating a market segment that is less efficient in terms of current pricing.

### 4.1.2 By-Class Statistical Metrics in OOS Predictions

We discuss the training and the testing by-class statistical metrics in this subsection. In addition to the exploitation of the uneven distribution of the return state transitions, our models attempt to balance the accuracy for each individual classes between the true positive predictions and the true negative predictions. Due to the specific purpose of our classification models, balancing between the true positive predictions and the true negative predictions is challenging. During the training process, our models create a one-versus-all multi-class classification structure, similar to one-hot encoding, and compare each individual return state against the other 9 return states. This introduces the unbalancedness. To the classification models, when looking at the data set with any one of the 10 return states, the data set will have a 90 % negative rate. By simply predictively label all the data points as negative, the model can achieve more than 90 % accuracy level for each single return state. However, if we denote negative with 0, we will be predicting all 0's across the 10 classes. In the end, we will have low overall accuracy and this adds no information at all. Therefore, balancing the true positive rate and the true negative rate becomes crucial for the success of our models.

Table 10 summarizes the key statistical metrics that measure the performance of our classification models in the OOS predictions covering 196301:201912 by class. With the detailed by-class metrics, we confirm that the models sacrifice the accuracy and the detection of the true positives for the return state 2, return state 3, and return state 4, regardless of the modeling architecture, potentially due to the unbalancedness of the data and the different level of uncertainties as discussed above. The prevalence column shows the ground truth distribution of the ten classes across the entire OOS testing period covering 196301:201912. As can be seen clearly, the return state 2, 3 and 4 are among the return states with the lowest number of observation in our entire sample. However, it is important to note that the return state 1 is also among the lowest populated states while the models collectively choose not to sacrifice the accuracy in return state 1. As reflected by the better accuracy across the models for return state 1, the models seem to spend more resources to improve the accuracy of prediction in return state 1. This again emphasizes that the uncertainty level of the return states are not equal. The extreme return states are more certain than the middle return states and the models are taking the advantage of this higher certainty. If we look at the summarizing measures that balance the accuracy between the true positives and the true negatives, the balanced accuracies for each of the return states are above 50 % indicating that there are gains beyond simply predictively labeling the individual return state as negative.

It is noteworthy that the proportion correctly predicted negative, Specificity is much higher than, the proportion of correctly predicted positives, Sensitivity. For Sensitively, the degree is extremely high for state 1, while the degree of Specificity is spread equally across states. This suggests the model is good at predicting the negative states rather than positive states,

## 4.2 Search for Predictable Individual Stocks and Measure the Models' Feelings

The existing literature utilizing the numeric prediction setup has not yet provided a picture of the predictable stocks. For the numeric prediction setups, it is challenging to come up with a good pre-realization measure about modeling certainty. Taking the advantage of our classification setup and the directly measured accuracy, we attempt to picture the stocks that are more predictable throughout their life in our sample at the individual stock level. We also further construct a measure of pre-realization modeling certainty to learn about the models' feelings about the predictions they are making. Note that this modeling certainty measure is a pre-realization measure, which means we can measure the models' feelings before looking at the ground truth return state in the next period.

### 4.2.1 Lifetime OOS Prediction Accuracy

To search for the predictable individual stocks, we first calculate the OOS prediction accuracy for each stock across time. Then, in each month, we normalize each characteristic across all stocks to obtain the relative positions of the characteristics for each stock in the month. We take the average of the normalized characteristics for each stock across time. Finally, we fit a regression model to examine the relation between the lifetime characteristics and the lifetime accuracy for all of the stocks. Table 11 presents the regression results based on the predictions made by GBM8 100, which is the best IS model in terms of accuracy. We categorize the characteristics into groups. Many characteristics are associated with the OOS prediction accuracy during the lifetime of individual stocks. Most of the characteristics have significant negative relation with the OOS prediction accuracy. However, some of the characteristics in the trading friction category are significantly positively associated with the OOS prediction accuracy.

Panel A in Table 13 presents the summarized results of the regression at the variable category level. We sum up all the regression coefficients of the characteristics that are significant at the 10% level. We select 10% as the threshold because we want to be more inclusive about the statistical effects from the different variable categories, and most of our variables are associated with very low p-values. Within Panel A in Table 13, it is clear that the individual stocks that have high trading friction are more predictable throughout their lifetime in the sample.

### 4.2.2 Lifetime Modeling Certainty

We further measure the model certainty level to learn about the models' feelings when making predictions. We specify the model certainty for a future return state prediction at a time point as the variance of predicted probabilities across all return states. Ideally, if a model is making a prediction that the model is sure about in a systematic way across the return states, the model should be able to distinguish the high probabilities from the low probabilities. Therefore, the variance of the predicted probabilities across all return states will be high. That is, the higher the model certainty, the higher the variance of the predicted probabilities. To conduct a regression analysis, for each stock, we first calculate the variance of the predicted probabilities for each prediction in each month. We then calculate the average of the variances across time for each stock and we use the average of variance as the response variable in our regression analysis. We follow the same procedure described before to produce individual stock level characteristics for lifetime across time. Finally, we regress the modeling certainty for each stock during its lifetime on the characteristics. Table 12 presents the regression results with the predictions made by GBM8 100. The characteristics are also grouped into variable categories. Most of the characteristics are significantly negatively associated with the modeling certainty. However, some trading frictions seem positively related to the model certainty.

Panel B in Table 13 reports the findings from Table 12 with variable categories. We sum up the coefficients associated with the characteristics that are significant in Table 12. Similar to the results in the case of OOS prediction accuracy, trading frictions are the only group of characteristics that has a positive relationship with the model certainty, i.e., models are more certain about the future return state predictions for the stocks that have higher trading frictions during their lifetime.

### 4.2.3 Discussion on Lifetime Prediction Accuracy

The findings documented in Section 4.2.1 is critical to the understanding of market efficiency at the individual stock level throughout the stocks' lifetime. The OOS predictability is a signal of the existence of historical information that has not been incorporated into prices. Following this line of thinking, at the individual stock level, trading frictions throughout the lifetime of the stocks seem the leading source for higher OOS predictability. In other words, because of the difference in trading frictions long the lifetime of the stocks, some stocks can be more predictable which is a signal of lower market efficiency. Because we are measure lifetime characteristics, this implies the possibility of the existence of segmented submarkets and stocks in different submarkets can experience different levels of market efficiency throughout their lifetime.

## 4.3 Biased Modeling Preference and the Adjustment to Improve Performance

In addition to the predictability throughout the lifetime of the stocks, we also study the possibility that the models have biased beliefs over certain return state outcomes. Predictions of the machine learning models are subjective. Therefore, models may have systematic bias in favoring certain outcomes when we classify the future return state based on the maximum predicted probability. An alternative to avoid this systematic bias in favoring certain outcomes is to use relative ranking of the predicted probabilities for each possible return state.

For example, a model can make many predictions to return state 6. If this is true, i.e., the models have systematically biased preference over some return state outcomes, it can deliver a reduced performance. To study this possibility, we look into the alternative way of using predicted probabilities to make OOS predictions. The default way of using the predicted probabilities is to only look at the maximum predicted probability and predict the future return state with the return state that is associated with the highest predicted probability. An alternative to the default method is to use the predicted probabilities relatively, i.e., we look at all of the 10 predicted probabilities and sort the 10 predicted probabilities into deciles. We then classify the future return states based on the deciles. Specifically, if a stock is in the top decile of a corresponding predicted probability, we can then classify the future return of the stock being in the return state associated with the decile. This gives us a way to form fuzzy portfolio allocation, i.e., stock can be in multiple return-state-based portfolios.

Specifically, we hold long positions in stocks with predicted return state 10 probability in the top decile of the month and we require the stocks to not be in the top deciles of the predicted return state probabilities for return state 1-9. Similarly, we short stocks with predicted return state 1 probability in the top decile of the month and we require the stocks to not be in the top deciles of the predicted return state probabilities for return state 2-10. At the same time, we adjust the weights of the stocks in the long and the short legs with their predicted return state 10 probabilities and their predicted return state 1 probabilities. In other words, among the selected stocks based on the relative predicted return state probabilities, we allocate more positions to stocks that have higher predicted probabilities of ending up in return state 1 and return state 10.

We also check the robustness of this modification, with an extremely conservative setup excluding all of the bottom 50% capitalization stocks and forming portfolios with value weights, the long-short portfolios are still able to deliver a Sharpe ratio as high as 3 folds of what the market can deliver. Table A.3 in the appendix presents the results.

## 5 Cross-Sectional Explanatory Power and Predictor Contribution

Next, we discuss the overall explanatory power of our models. We construct a separate sample to test the model structures cross-sectionally. We demonstrate the average rank of variable importance across the two types of modeling architectures, i.e., the neuron networks and the tree models. We connect the predictor contribution to the good performance of our models in the cross-sectional OOS evaluation and further discuss the new insights about EMH based on the variable contributions.

### 5.1 Cross-Sectional Explanatory Power of the Models[6]

We conduct cross-sectional (CS) tests to examine whether our models can capture the overall return state changes. To do this, we split the CRSP-COMPUSTAT universe into two subsamples including odd number months and even number months in the spirit of Fama and French (2018). We use odd number months as our training sample and make CS OOS tests with even number months where the observations are new to the models. This setup for the IS training and the OOS testing enables us to directly look at the overall explanatory performance of our models across the entire coverage of CRSP-COMPUSTAT universe from 1963 to 2019.

Figures 12 and 13 show the CS OOS performance with economic metrics calculated based on the testing set of our CS sample splitting, i.e., the even number months that the models have never seen during the IS training process with the odd number months. The performance in CS OSS tests is similar to what we observed in the TS OOS tests. First, our classification based long-short portfolios systematically outperform the market portfolio. Second, we observe a similar increasing trend of performance as we adjust the complexity of the models. Third, the best performing models in terms

[6] We are updating the results from cross-sectional setup.

of the CS OOS economic metrics match the best performing models in TS OOS tests. We see that the single-layer neuron network with 128 neurons, ANN1 128, performs the best among the neuron networks, and GBM8 100 and DART8 100 perform the best among the tree models.

We also include a table of the overall accuracy of the classification. We can see in Table 17 that our model prediction accuracy levels are still higher than the no information accuracy in CS OOS predictions. This means that our models overall obtained information about the OOS return transitions based on looking at the IS observations. In other words, our models can explain the relative performance of the stock in a traditional cross-sectional setup to the degree of accuracy that is statistically meaningful.

### 5.2 Variable Importance across Models

With the good IS and OOS performance demonstrated through both the TS setup and the CS setup, we further look into the predictor contributions measured as variable importance across models and discuss what are the driving predictors that lead to the superior performance of our models across the different setups. Table 8 summarizes the variable importance with the average values, including the rank of the contribution to each model. We separately demonstrate the variable importance for the neuron networks and the tree models, since the models have structural differences in dealing with categorical inputs. We discuss the variable importance of the CS training models presented in Table 18 and include the variable importance of the TS training models in the appendix.[7].

Unlike what is documented by Gu et al. (2020) or Chen et al. (2020), the top contributor in both the neuron networks and the tree models is the idiosyncratic volatility. The monthly average of the daily bid-ask spread divided by the average of daily spread makes the second important contribution across the modeling architectures. It is worth mentioning that the tree models seem highly dependent on idiosyncratic volatility, bid-ask spread, and return volatility, while the contribution made by various top contributing predictors in neuron networks are more balanced. Beyond the top contributors, across the modeling architectures, all types of historical information make an important contribution. Specifically, we see that other trading-related variables and the industry indicators make a huge portion of the contribution. At the same time, the historical corporate announcements, such as earning price ratio, IPO status, convertible debt obligation, firm R&D expenses, change of the number of analysts, etc., all make a substantial contribution to our models. Macroeconomic indicators are also among the top 50 contributing predictors. However, the contribution of macroeconomic data components is very limited.

The fact that the historical information, including return related information, corporate announcements, can contribute to the model predictability is interesting to the understanding of the market efficiency. The semi-strong form of market efficiency permits the generation of excess profit through information asymmetry and lowers the bar to focus on the speed at which the public information is

[7] The table will be included with the next update

incorporated into the prices. However, the contribution from corporate announcements coupled with the strong predictability across the setups shows that there exists systematic relation between the future returns and the lagged corporate announcement variables. Considering that some of the variables, such as R&D are lagged by at least 6 months, the semi-strong form of market efficiency seems questionable. If the market is efficient in the semi-strong form, it is hard to explain why the corporate announcements from 6 months ago can still help predict the future return states.

In addition, the weak form of market efficiency states that the historical prices and trends cannot predict future returns. Yet, the top-ranked contributors to our models are populated with the past return and trading information. In fact, while the contribution from corporate announcements and the contribution from the past return information have ratio of 50% : 50% in the tree models, the past return information makes more contribution to our neuron network models comparing to the corporate announcements. This shows that the historical return information can help predict the future return states for the individual stocks and thus questions the weak form of market efficiency.

## 6 Conclusion

In this paper, we introduce the machine learning classification methods to the asset pricing literature and examine market efficiency. Taking the advantage of the relation between classification and information theory, we force the models to extract the information about the relation between the historical information corresponding to the different forms of market efficiency.

We analyze the economic performance of our classification portfolios in terms of OOS return distribution, SR, CEQ, and maximum drawdown. Our classification-based portfolios beat the market systematically in multiple setups. The best models demonstrate surprisingly good performance across the metrics. We also measure the OOS performance with statistical metrics. We see that the OOS prediction accuracies match the OOS economic performance of the associated classification portfolios. We take the advantage and utilize the accuracy metric, which is only applicable to classification problems, as a proxy to study whether the future returns of individual stocks are independent of multiple types of historical information. We introduce the binomial test to the asset pricing literature and document the statistical significance of the classification model accuracy against the no information accuracy.

Our findings of statistical significance have important implications. First, the accuracy indicates that there is a meaningful relationship between the future return states and the lagged predictors representing historical information. In other words, the prediction accuracy is statistically meaningful and the future return states are predictable. This is the first time in the literature that the market efficiency is examined through the prediction accuracy as the proxy. Second, our classification models successfully capture the relation between the historical information and the future return states in a predictive format, indicating that the new information beyond the distribution of the return states has been generated with our classification models. This demonstrates the possibility that sophisticated investors can apply complex tools, such as the machine learning classification methods, to generate new information that is not reflected by the market prices. Specifically, the generation of new information about future return states by the sophisticated investors is equivalent to manually introduce the information asymmetry to the market. The sophisticated investors can take the information advantage against the other investors and benefit from it in their trading activities.

We also document important economic insights that are unique to the literature. First, we demonstrate findings on the transitions of the return states. The ground truth returns state transitions during the period of 196301:201912 show uneven levels of uncertainty. The extreme state-related transitions are with substantially higher certainty comparing to the other transitions. We show that our models learn about this difference of uncertainty and take the advantage of it. The accuracy by return state transitions and the uneven uncertainty we demonstrate collectively imply that different return state transitions may be related to different levels of market efficiency. Second, at the individual stock level, we identify the lifetime characteristics that are associated with the lifetime OOS predictability. We show that the trading frictions are the only source that is positively related to the OOS predictability for individual stocks. This finding is critical. In the market, there may be segmented submarkets with different levels of trading frictions. For stocks in different segmented submarkets during their lifetime, they can experience different levels of market efficiency as signaled by different levels of OOS predictability. Third, machine learning models may also have biased preferences over certain predicted outcomes. Because of this type of bias, the OOS performance of the portfolios can decrease. Correcting the biased preference of the models can increase the OOS performance systematically.

In terms of the contribution of the predictors, we show that historical information including return-related information, corporate announcements, macroeconomic indicators, etc. all make an important contribution to our models, despite that the contribution from macroeconomic data components is very limited. The fact that the historical information including return information and corporate announcements being able to make a contribution to predictability challenges the weak form of market efficiency and the semi-strong form of market efficiency. Combined with the implication that the sophisticated investors may be able to generate new information based on historical information, we conclude that there is still room for the market efficiency to improve.

# References

Barberis and Thaler, 2003, A Survey of Behavioral Finance, ch. 18, p. 1053-1128 in Constantinides, Harris and Stulz eds., *Handbook of the Economics of Finance*, vol. 1, Part 2, Elsevier.

Bianchi, Daniele, Buchner, Mathhias and Tamoni, Andrea, 2020, Bond Risk Premia with Machine Learning, Queen Mary University of London, University of Warwick and Rutgers working paper.

Brandt, Michael, Santa-Clara, Pedro, and Valkanov, Rossen, 2007, Parametric Portfolio Policies: Exploiting Characteristics in the Cross-Section of Equity Returns, *Review of Financial Studies* 22(9), 3411-3447.

Bryzgalova, Svethana, Pelger, Markus, and Zhu, Jason, 2020, Forest Through the Trees: Building Cross-Sections of Stock Returns, London Business School and Stanford University working paper.

Carhart, Mark M., 1997, On Persistence in Mutual Fund Performance, *Journal of Finance* 52(1), 57-82.

Chen, Luyang, Pelger, Markus and Zhu, Jason, 2020, Deep Learning in Asset Pricing, Stanford University working paper.

Cohen, Malloy and Nguyen, 2020, Lazy Prices, *Journal of Finance* 75(3), 1371-1415

DeMiguel, Victor, Garlappi, Lorenzo, and Uppal, Rama, 2009, Optimal Versus NaÃ¯ve Diversification: How Inefficient is the 1/N Portfolio Strategy? *Review of Financial Studies* 22(5), 1915-1953.

Fama, 1969 (May, 1970), Efficient Capital Markets: A Review of Theory and Empirical Work, *Journal of Finance*, 25(2), Papers and Proceedings of the Twenty-Eighth Annual Meeting of the American Finance Association New York, N.Y. December, 28-30, 383-417.

Fama, 1991, Efficient Capital Markets: II, *Journal of Finance*, 46, 1575-1617.

Fama, Eugene, and French, Fama, 1992, The Cross-Section of Expected Stock Returns, *Journal of Finance* 47(2), 427-465.

Fama, Eugene, and French, Fama, 2018, Choosing Factors, *Journal of Financial Economics* 128(2), 234-252.

Feng, Guanhao, Polson, Nicholas G., Xu, Jianeng, 2019, Deep Learning in Characteristics-Sorted Factor Models, University of Chicago working paper.

Goyal, Amit, and Welch, Ivo, 2008, A Comprehensive Look at the Empirical Performance of Equity Premium Prediction, *Review of Financial Studies* 21(4), 1455-1508.

Green, Jeremiah, Hand, John and Zhang, Frank, 2017, The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns, *Review of Financial Studies* 30(12), 4389-4436.

Grossman and Stiglitz, 1980, *American Economic Review*, 70(3), 393-408.

Gu, Shihao, Kelly, Bryan, and Xiu, Dacheng, 2020, Empirical Asset Pricing via Machine Learning, *Review of Financial Studies* 33(5), 2223-2273.

Hou, Kewei, Mo, Haitao, Xue, Chen and Zhang, Lu, 2019, Which Factors? *Review of Finance* 23(1), 1-35.

Hou, Kewei, Xue, Chen, and Zhang, Lu, 2015, Digesting Anomalies: An Investment Approach, *Review of Financial Studies* 28(3), 650-705.

McCracken, Michael, and Ng, Serena, 2016, FRED-MD: A Monthly Database for Macroeconomic Research, *Journal of Business & Economic Statistics* 34(4), 574-589.

Moritz, Benjamin and Zimmermann, Tom, 2016, Tree-Based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns, Ludwig Maximilian University of Munich and University of Cologne working paper.

Rau, 2011, Market Inefficiency, ch. 18, p. 331-349 in Baker and Nofsinger eds., *Behavioral Finance: Investors, Corporations and Markets*, Wiley.

Rossi, Alberto, 2018, Predicting stock market returns with machine learning, Georgetown University working paper.

Sargent, 1994, *Bounded Rationality in Macroeconomics*, A Clarendon Press Publication.

Shannon, Claude, 1948, A Mathematical Theory of Communication, *Bell System Technical Journal* 27(3), 379-423.

Wolff, Dominik, and Echterling, Fabian, 2020, Stock Picking with Machine Learning, Darmstadt University of Technology and Deka Investment GmbH working paper.

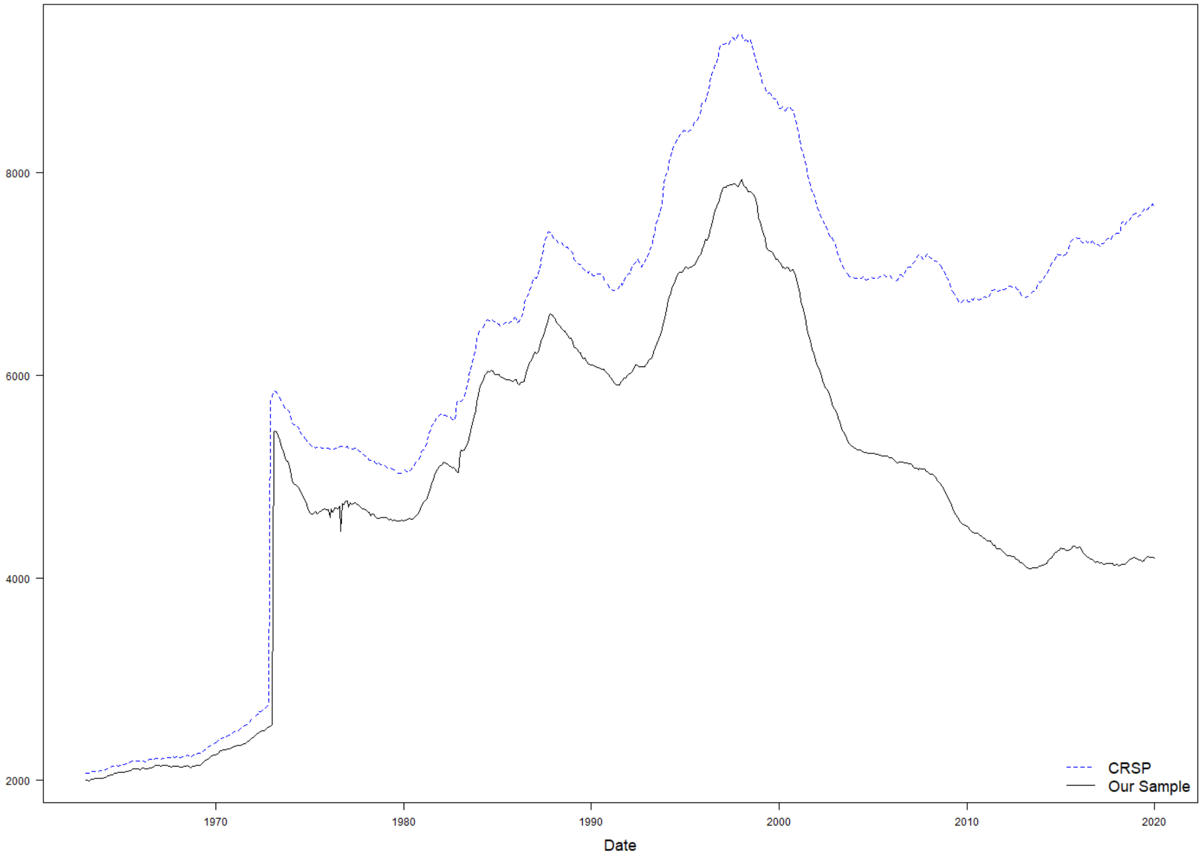

**Figure 1 Number of Stocks in CRSP vs Number of Stocks in Our Sample 196301:201912**

Figure 1 presents a comparison of the sample coverage between our data set and the CRSP database. The dashed line represents the number of securities included in the CRSP database and the solid line represents the number of stocks included in our sample. Note that CRSP is a general security database. It includes securities other than stocks of the public firms. In our sample, we include only the stocks listed on NYSE, Amex, and NASDAQ. This figure presents the comparison from January 1963 to December 2019. In total, our sample covers distinct 26302 stocks. On average, our sample covers around 4887 stocks for every trading month. The detailed summary statistics of the sample coverage can be found in Table 4.

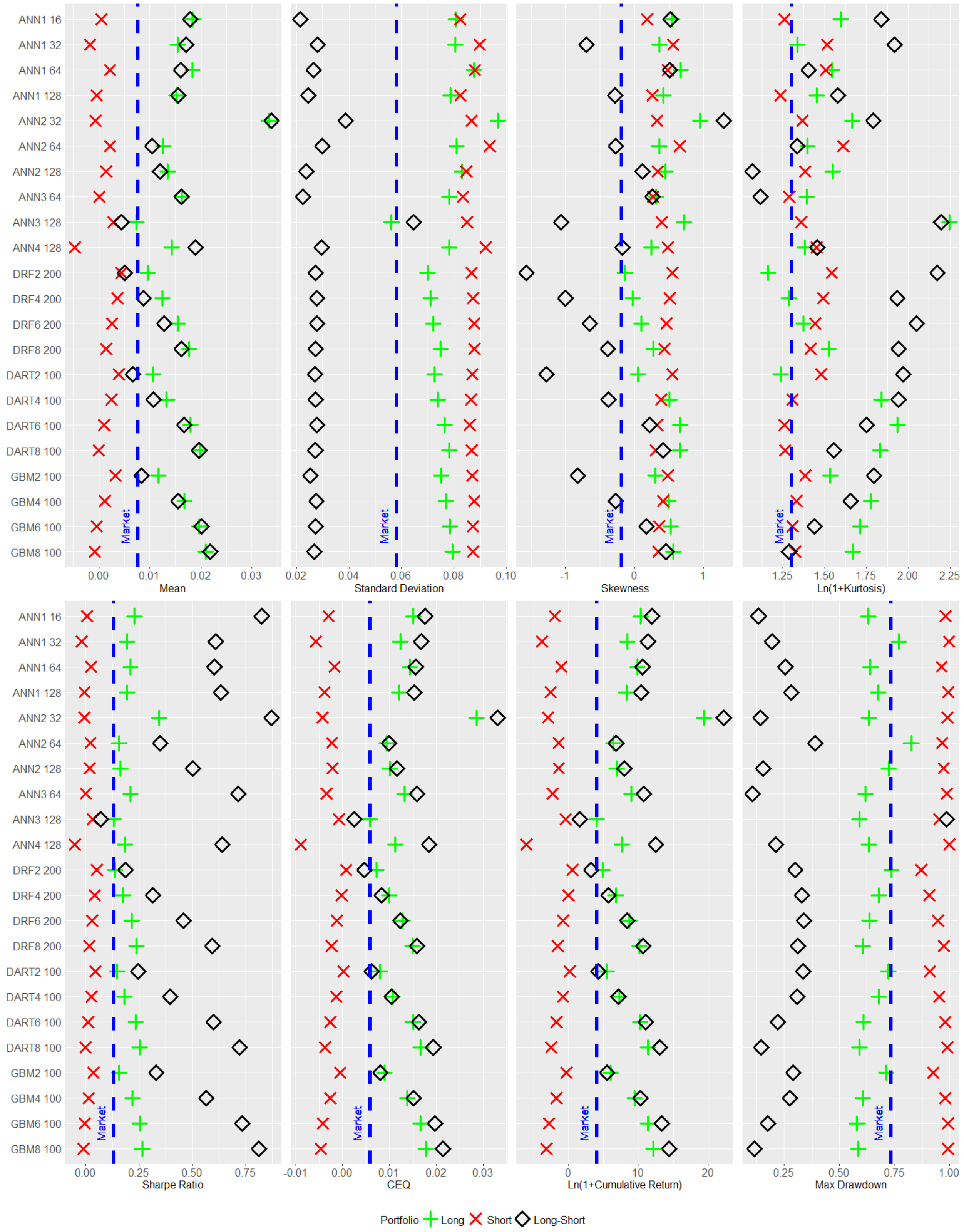

**Figure 2 Equal Weight Time Series OOS Portfolio Economic Performance 196301:201912**

**Figure 2 (Continues)**

Figure 2 summarizes OOS economic metrics of the classification-based portfolios covering 196301:201912 with the equal-weight scheme across different classification models. More specifically, a long portfolio represented by "+" indicates *a long position* in all stocks predicted to be in the return state 10, or the best return state. A short portfolio represented by "X" indicates *a long position* in all stocks predicted to be in the return state 1, or the worst return state. A long-short portfolio represented by "◊" indicates *a mixed portfolio including a long position in the return of all stocks predicted in the return state 10 and a short position in all stocks predicted in the return state 1*. In each of the subgraphs, the dashed line indicates the reference performance delivered by the market portfolio, i.e. buy-hold strategy applied to the entire market. The OOS performance of the stocks is equal weighted. The allocation to the long leg and the short leg in a long-short portfolio is always equal, i.e., having a 50%:50% allocation ratio. All the returns are fully risk-adjusted against risk free rate. We measure the plain monthly Sharpe Ratio (SR) as return scaled by standard deviation. To convert our SR to annualized SR, one needs to multiply our SR with square root of 12. We follow the literature and adopt $\gamma = 1$ in the calculation for CEQ. The cumulative returns are calculated as the gross returns net of the initial investment. Details about the economic metrics are discussed in Section 2.2.1 and the model specifications are presented in Table 2.

The equal weight long-short portfolios based on our classification models systematically outperform the market. For the tree models, as the modeling complexity increases, the performance of our models increases. Our equal weight long-short portfolios deliver surprisingly good performance in term of controlling left-tail risk while delivering the amazing level of economic performance. Specifically, from 196301:201912, the equal weight long-short portfolio based on our tree model GBM8 100 delivers a SR of 0.81 and the equal weight long-short portfolio based on our ANN model ANN2 32 delivers a SR of 0.87. This performance is comparable to the plain monthly SR of 0.707 achieved by the best equal weight portfolio in Gu, Kelly and Xiu (2020) as in their appendix table A.9 (their OOS data, or testing sample, covers from 1987:2016 and they have year-by-year rolling model updating). The maximum drawdowns associated with the equal weight long-short portfolios of ANN2 32 and GBM8 100 are at 14.23% and 11.57 % respectively. During the same time period, the equal weight market portfolio achieves a SR of 0.1308 and a maximum drawdown of 73.14 %.

Note that Figure 2 is based on our models using only firm characteristics and the industry information. According to our comparison, modeling with other components, including characteristics-based factor mimicking returns, Goyal and Welch (2009) predictors and macroeconomic indicators provided by FRED-MD, does not change the performance much. Table 6 provides a related comparison. These macro-level components make limited contribution to the accuracy as measured by variable importance. Table 15 presents the variable importance with our cross-sectional models including all data components. Our Figures 2 to 11 are all based on our data without macro-level data components.

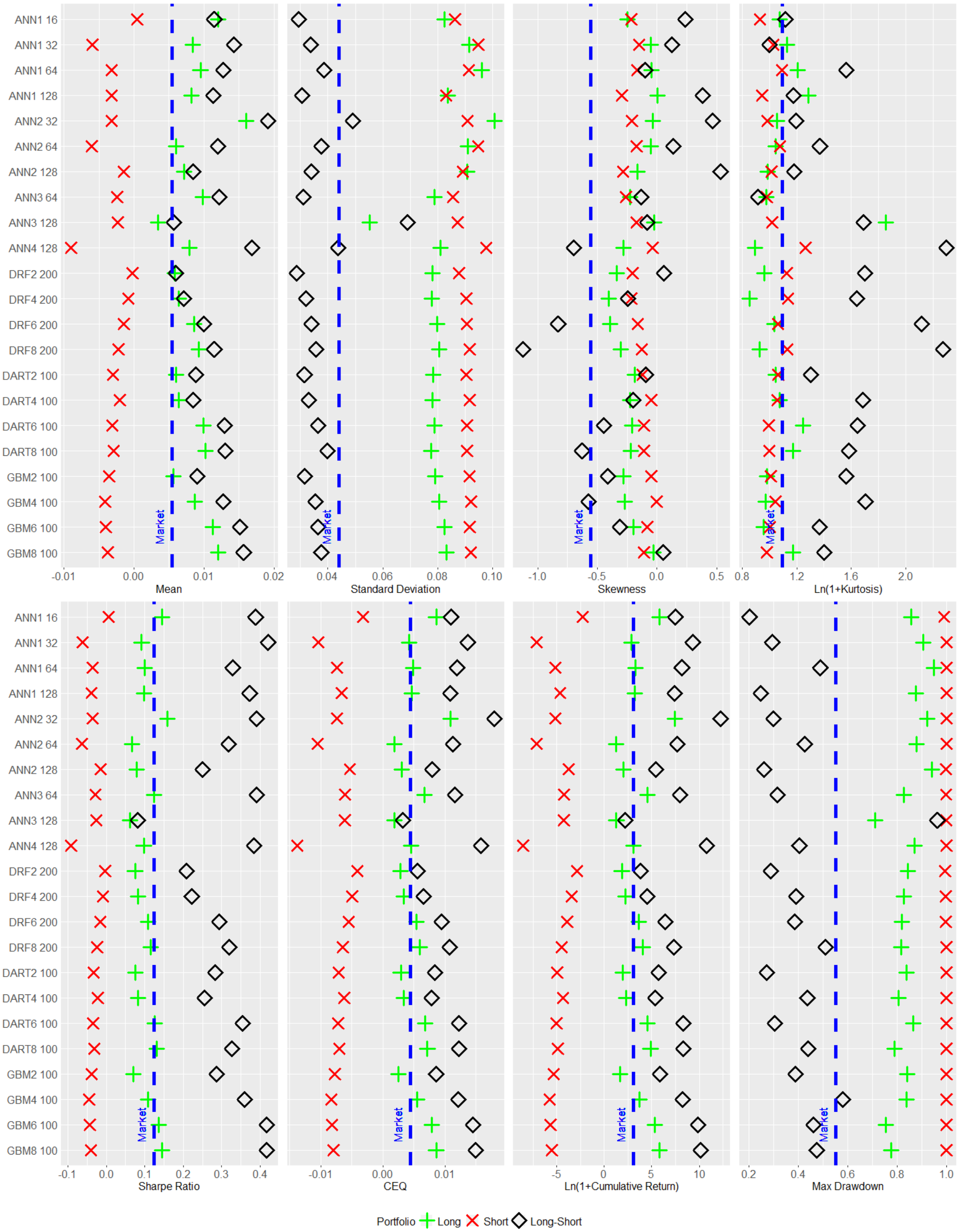

**Figure 3 Value Weight Time Series OOS Portfolio Economic Performance 196301:201912**

**Figure 3 (Continues)**

Similar to Figure 2, Figure 3.summarizes OOS economic metrics of the classification-based portfolios covering 196301:201912 with the value-weight scheme across different classification models. In general, the change of weight scheme does not affect the OOS performance of our classification portfolios.

Specifically, a long portfolio represented by "+" indicates *a long position* in all stocks predicted to be in the return state 10, or the best return state. A short portfolio represented by "X" indicates *a long position* in all stocks predicted to be in the return state 1, or the worst return state. A long-short portfolio represented by "◊" indicates *a mixed portfolio including a long position in the return of all stocks predicted in the return state 10 and a short position in all stocks predicted in the return state 1*. In each of the subgraphs, the dashed line indicates the reference performance delivered by the market portfolio, i.e. buy-hold strategy applied to the entire market. The OOS performance of the stocks is value weighted. The allocation to the long leg and the short leg in a long-short portfolio is always equal, i.e., having a 50%:50% allocation ratio. All the returns are fully risk-adjusted against risk free rate. We measure the plain SR as return scaled by standard deviation. To convert our SR to annualized SR, one needs to multiply our SR with square root of 12. We follow the literature and adopt $\gamma = 1$ in the calculation for CEQ. The cumulative returns are calculated as the gross returns net of the initial investment. Details about the economic metrics are discussed in Section 2.2.1 and the model specifications are presented in Table 2.

Similar to the equal weight version of the OOS tests in economic metrics, the value weight long-short portfolios based on our classification models systematically outperform the market. As the modeling complexity increases, the performance of our models increases. Our value weight long-short portfolios also deliver good performance in term of controlling left-tail risk while delivering the amazing level of economic performance. Specifically, for the OOS period from 196301:201912, our value weight long-short portfolios based on our models, ANN1 32, GBM8 100, GBM6 100, deliver OOS SRs of 0.4214, 0.418 and 0.418, comparable to the best plain monthly SR of 0.39 achieved by Gu, Kelly and Xiu (2020) as demonstrated in their Table 7. The maximum drawdown of the long-short portfolios based on ANN1 32, GBM8 100 and GBM6 100 are 29.47 %, 47.48% and 46.35%. During the same time period, the value weight market portfolio achieves a SR of 0.123 and a maximum drawdown of 55.1 %.

Note that Figure 3 is based on our models using only firm characteristics and the industry information. According to our comparison, modeling with other components, including characteristics-based factor mimicking returns, Goyal and Welch (2009) predictors and macroeconomic indicators provided by FRED-MD, does not change the performance much. Table 6 provides a related comparison. These macro-level components make limited contribution to the accuracy as measured by variable importance. Table 15 presents the variable importance with our cross-sectional models including all data components. Our Figures 2 to 11 are all based on our data without macro-level data components.

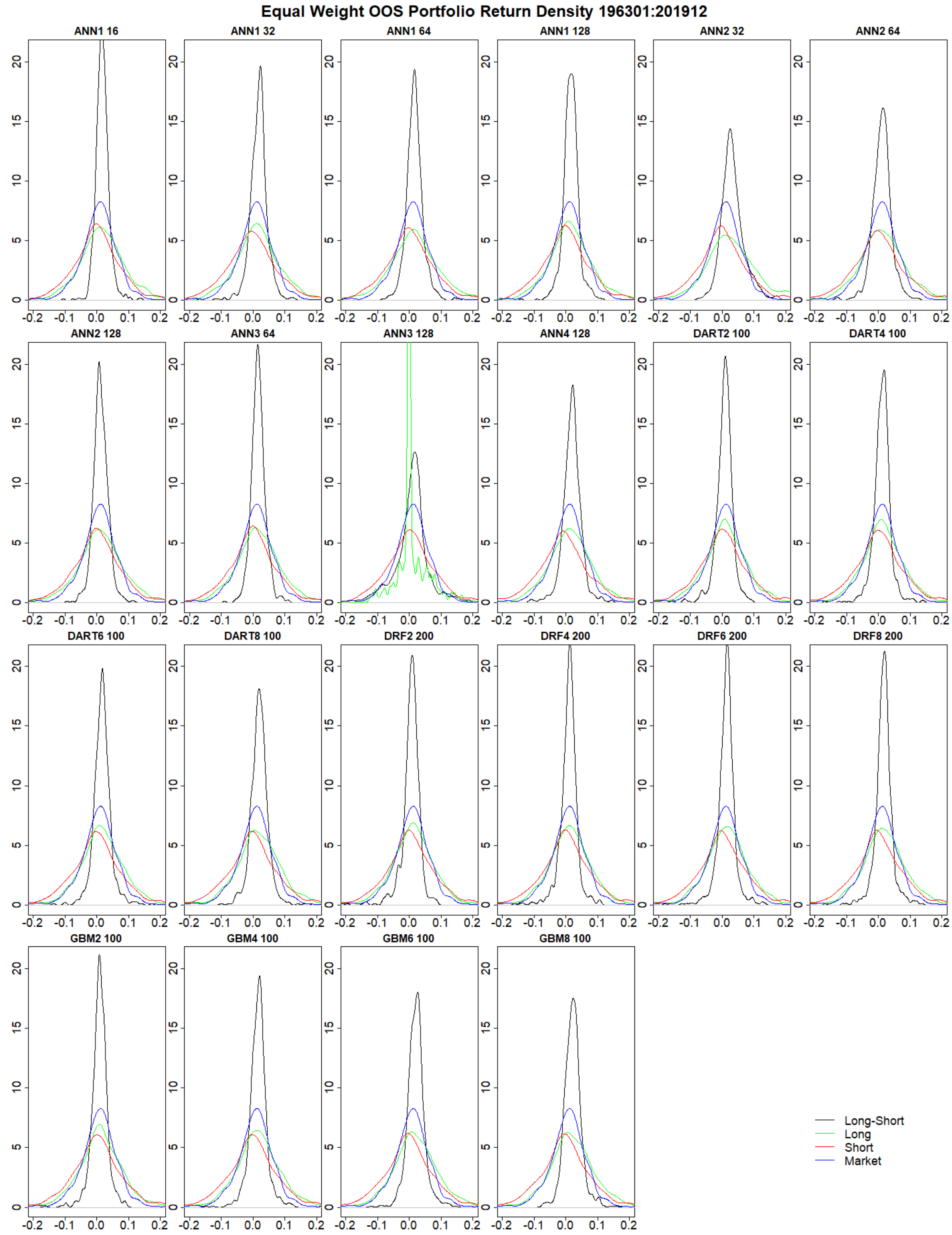

**Figure 4 Equal Weight OOS Portfolio Return Distributions 196301:201912**

**Figure 4 (Continues)**

Figure 4 presents the overlaid comparison of the OOS portfolio return distributions across the different allocation strategies based on different classification models. The lines in blue, green, red and black represent respectively the OOS return distributions of the market portfolio, the equal weight long position in the predicted lowest return state (labeled as "short" in the figure), the equal weight long position in the predicted highest return state (labeled as "long" in the figure), and the equal weight long-short position. Overall, the gaps between the green lines and the red lines indicate that our models are able to distinguish the worst OOS performing stocks and the best OOS performing stocks from each other. Our equal weight long-short portfolios provide overall return distributions with a concentration shifted towards the right tails and the variances are significantly reduced. The return distributions included in Figure 4 are all from the equal weight portfolios with the similar construction procedure mentioned in Figure 2.

The best OOS equal weight long-short portfolio monthly average return in our test is delivered by our neural network model ANN2 32. The long-short portfolio delivers an average of 3.4 % monthly return from 196301:201912. The long-short portfolio based on our tree model, GBM8 100, delivers an average monthly return of 2.2 %. During the same period, the market portfolio delivers an average monthly return of 1.1 %.

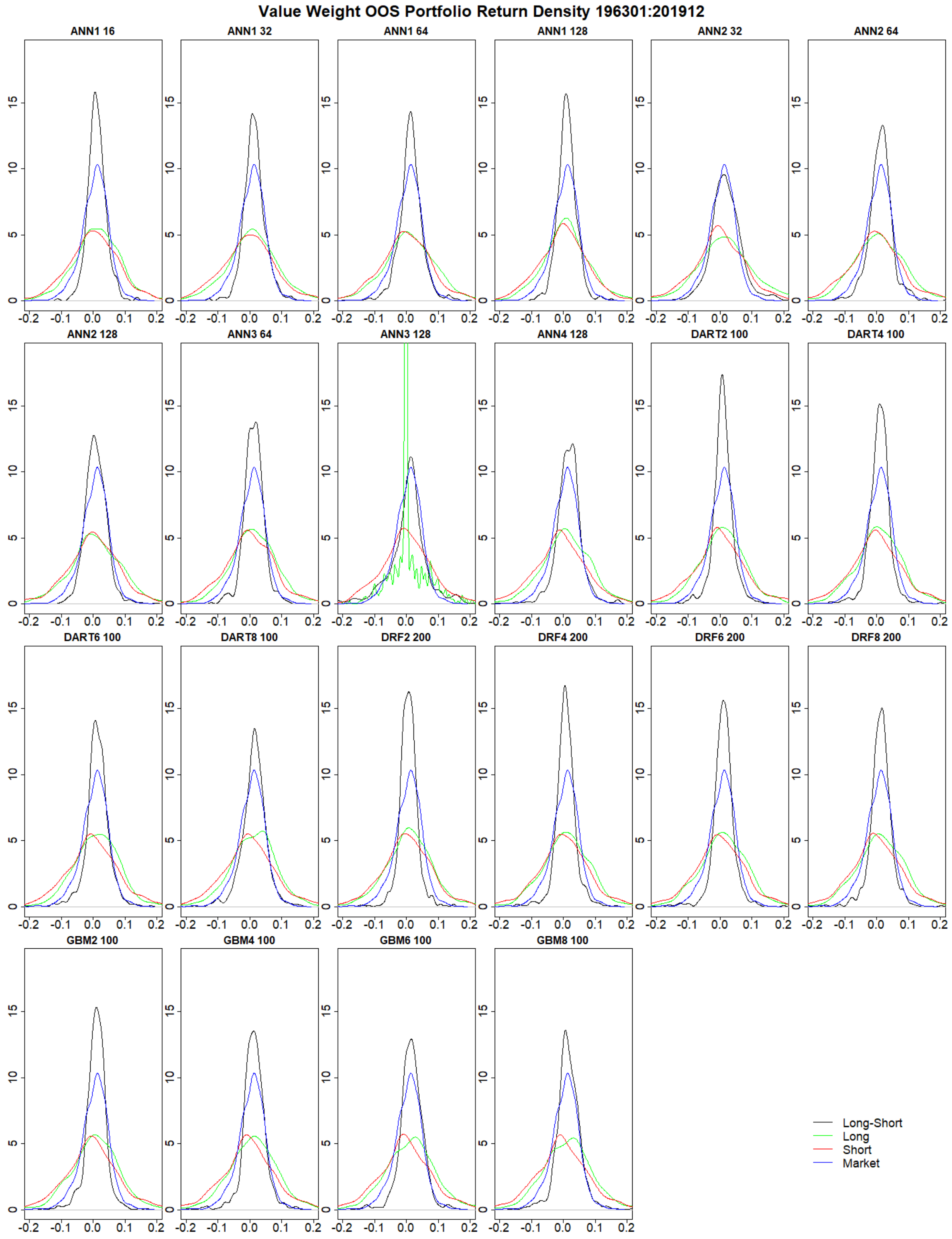

**Figure 5 Value Weight OOS Portfolio Return Distributions 196301:201912**

**Figure 5 (Continues)**

Figure 5 presents the return distribution comparison similar to Figure 4 but of value weight portfolios. The lines in blue, green, red and black represent respectively the OOS return distributions of the market portfolio, the value weight long position in the predicted lowest return state (labeled as “short” in the figure), the value weight long position in the predicted highest return state (labeled as “long” in the figure), and the value weight long-short position. Similar to what is presented in Figure 4, the gaps between the green lines and the red lines indicate that our models are able to distinguish the worst OOS performing stocks and the best OOS performing stocks from each other. Our value weight long-short portfolios also provide overall return distributions with a concentration shifted towards the right tails and the variances are significantly reduced. The return distributions included in Figure 5 are all from the value weight portfolios with the similar construction procedure mentioned in Figure 3.

For the value weight scheme, the long-short portfolio based on our neural network model ANN2 32 delivers an OOS average monthly return of 1.92 % from 196301:201912, while the market portfolio delivers an average monthly return of 0.9 %.

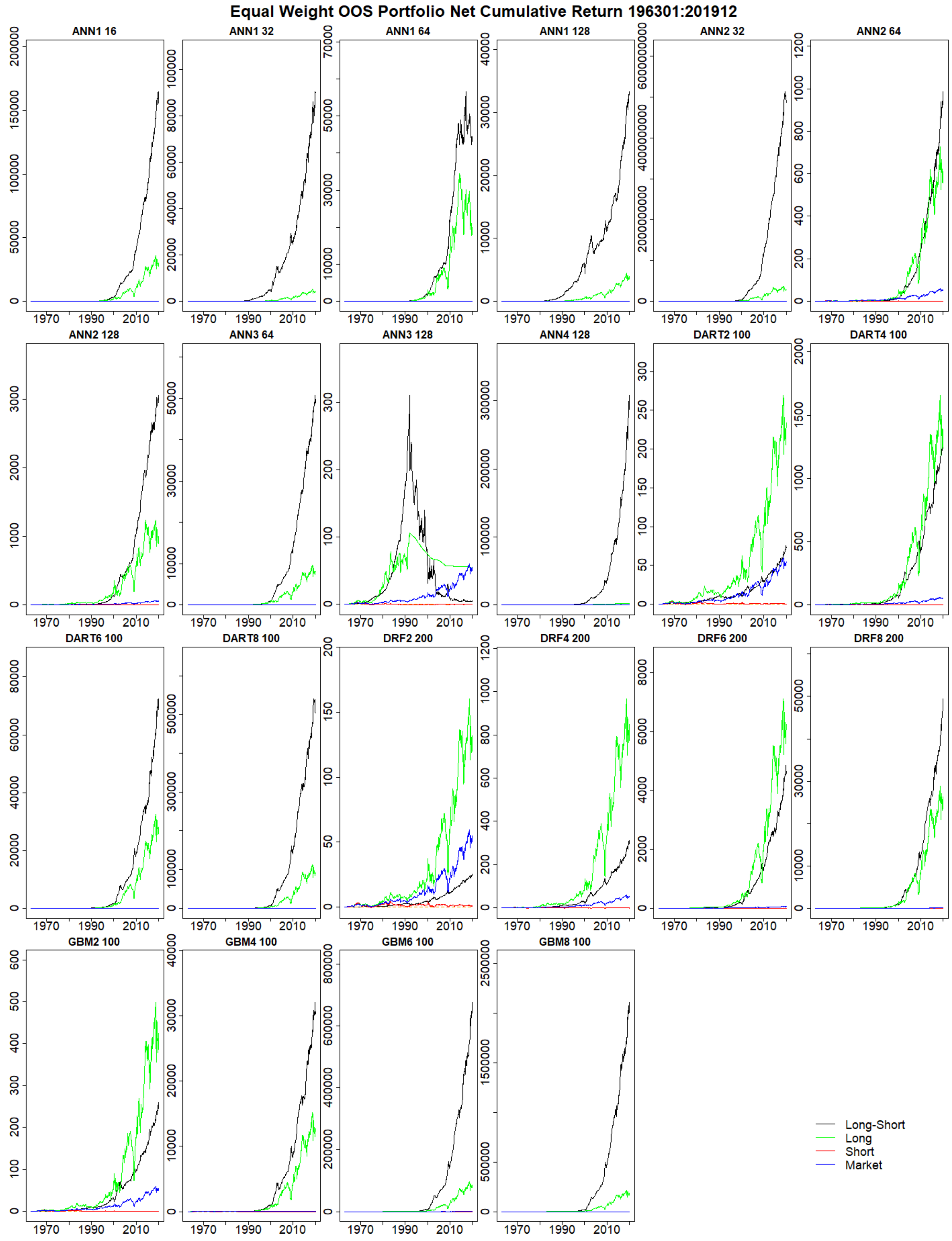

**Figure 6 Equal Weight OOS Portfolio Cumulative Returns 196301:201912**

**Figure 6 (Continues)**

Figure 6 shows the cumulative returns (in 100%) of our equal weight portfolios based on different classification models. The lines in blue, green, red and black represent respectively the OOS cumulative returns of the market portfolio, the equal weight long position in the predicted lowest return state (labeled as “short” in the figure), the equal weight long position in the predicted highest return state (labeled as “long” in the figure), and the equal weight long-short position.

Our equal weight long-short portfolios deliver phenomenal cumulative returns in the OOS test. Comparing to our long-short portfolios, the market portfolios is a horizontal line in the subgraphs as the cumulative return delivered by the market over the same period is too small. Our Neural Network model, ANN2 32 OOS delivers a cumulative return of 485,649,224,178.58%. The long leg of ANN2 32 alone delivers a cumulative return of 29,138,823,156.42% from 196301:201912. The equal weight long-short portfolios based on our tree models, GBM8 100 and GBM6 100, deliver OOS cumulative returns of 200,202,148.41% and 64,231,407.4% respectively during the same investment period. During the same period, the market portfolio achieves a cumulative return of 72,874%, substantially lower than what is delivered by our tree model DRF4 200. Despite that DRF4 200 is clearly underfitted, it still delivers a cumulative return of 84,804.67%.

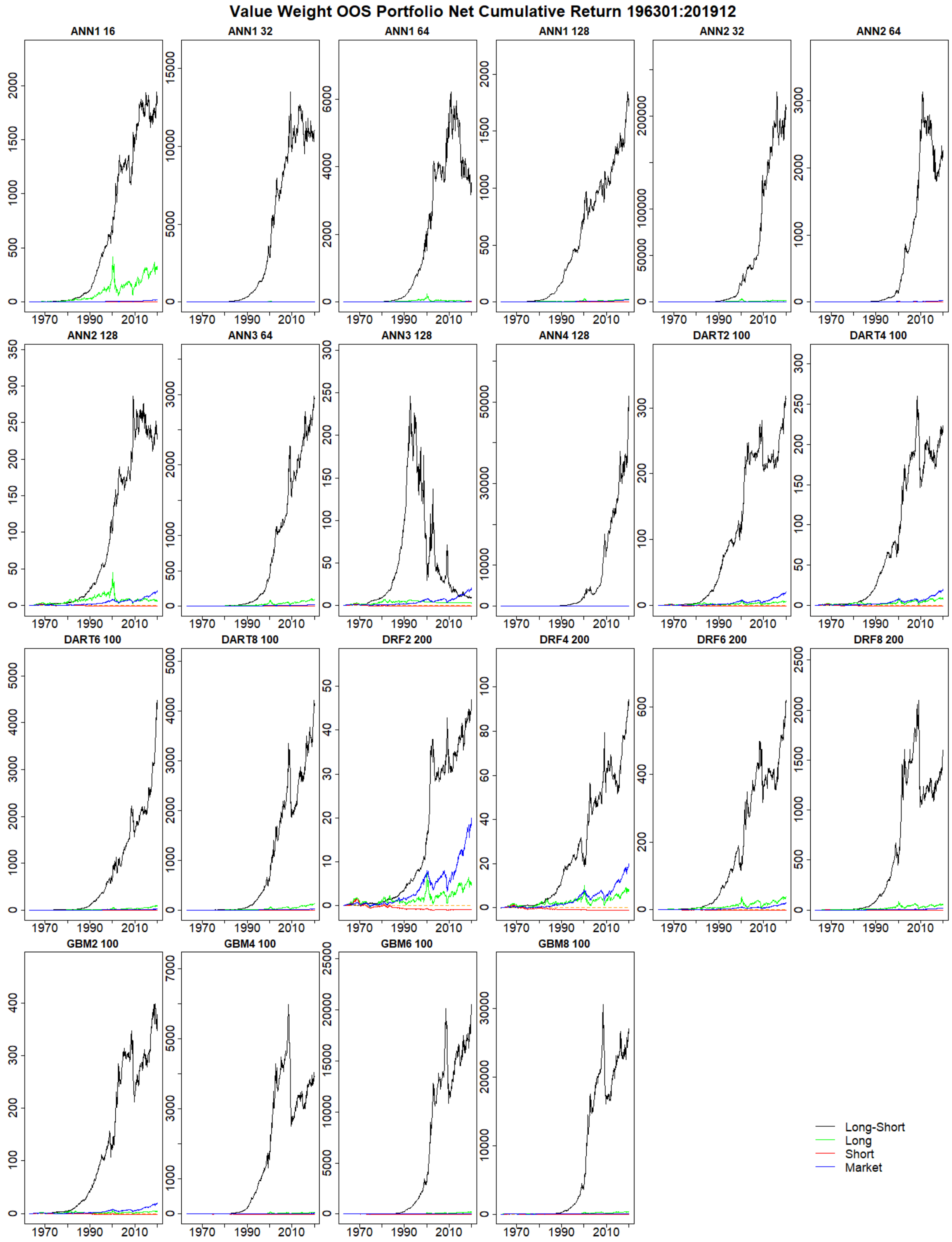

**Figure 7 Value Weight OOS Portfolio Cumulative Returns 196301:201912**

**Figure 7 (Continues)**

Figure 7 is the counterpart of Figure 6 for value weight portfolios. The lines in blue, green, red and black represent respectively the OOS cumulative returns of the market portfolio, the value weight long position in the predicted lowest return state (labeled as "short" in the figure), the value weight long position in the predicted highest return state (labeled as "long" in the figure), and the value weight long-short position.

Our portfolios in value weight also deliver shocking OOS cumulative returns comparing to what the market is capable to achieve. Our long-short portfolios based on neural network models, ANN2 32, ANN4 128 and ANN1 32, achieve OOS cumulative returns of 20,738,715.94 %, 4,784,930.78 % and 1,103,697.15% respectively from 196301:201912. The long-short portfolios based on our tree models, GBM8 100, GBM6 100 and DART6 100, deliver cumulative returns of 2,653,376.44%, 1,988,854.31% and 429,354.55% respectively. During the same period, the value weight market portfolio deliver a cumulative return of 27,396 % less than the cumulative return delivered by the long-short portfolio based on our tree model DART2 100.

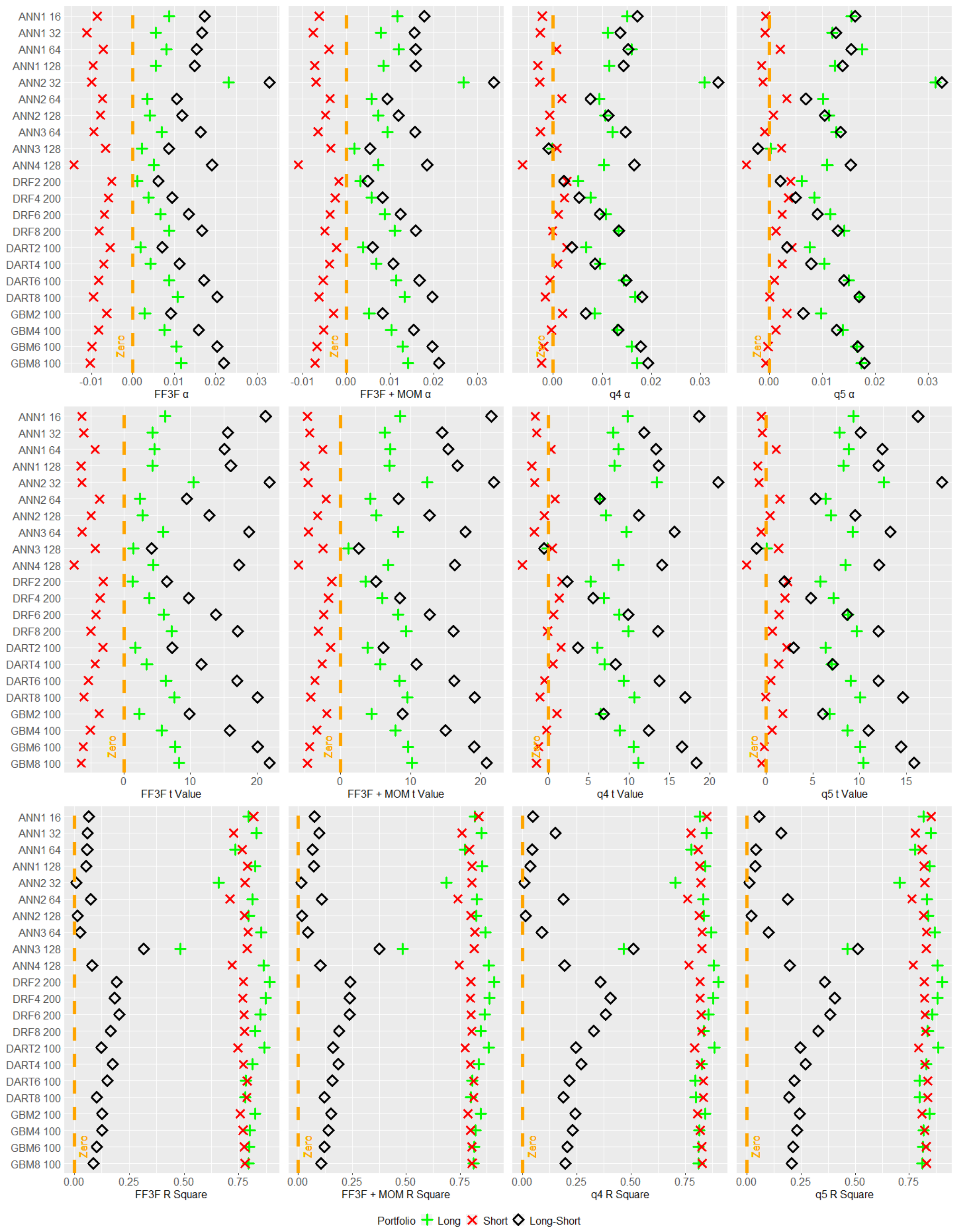

**Figure 8 Factor Model Tests on Equal Weight OOS Portfolio Returns 196301:201912**

**Figure 8 (Continues)**

Figure 8 demonstrates the factor model tests for the equal weight portfolios. The legends are the same as in Figure 2 and 3 except that the orange dashed line stands for 0. We obtain factor data sets from Kenneth French's data library and Lu Zhang's website. FF3F stands for the Fama French 5 Factor model. MOM stands for the momentum factor. q4 model is the investment CAPM from Hou, Xue, Zhang (2015) and q5 model is the update of the q4 model including the expected growth factor (R_EG). Note that we do not include Fama French 5 Factor model as we include q4 and we do not include MOM to any of the q-factor models because the q-factor models explains MOM. We present the model alphas, the t statistics of the model alphas and the R square of the models in the 3 lines of subgraphs. It is obvious that the factor models cannot explain the returns achieved by our long-short portfolios. For example, our long-short portfolio based on our neural network model, ANN2 32, has alphas of 3.31 %, 3.37 %, 3.35 % and 3.26 % respectively against FF3F, FF3F + MOM, q4 and q5. All R squares of the associated models are below 3 %. The explanatory power of the factor models on the portfolios seems decreasing as the complexity of our classification models increases.

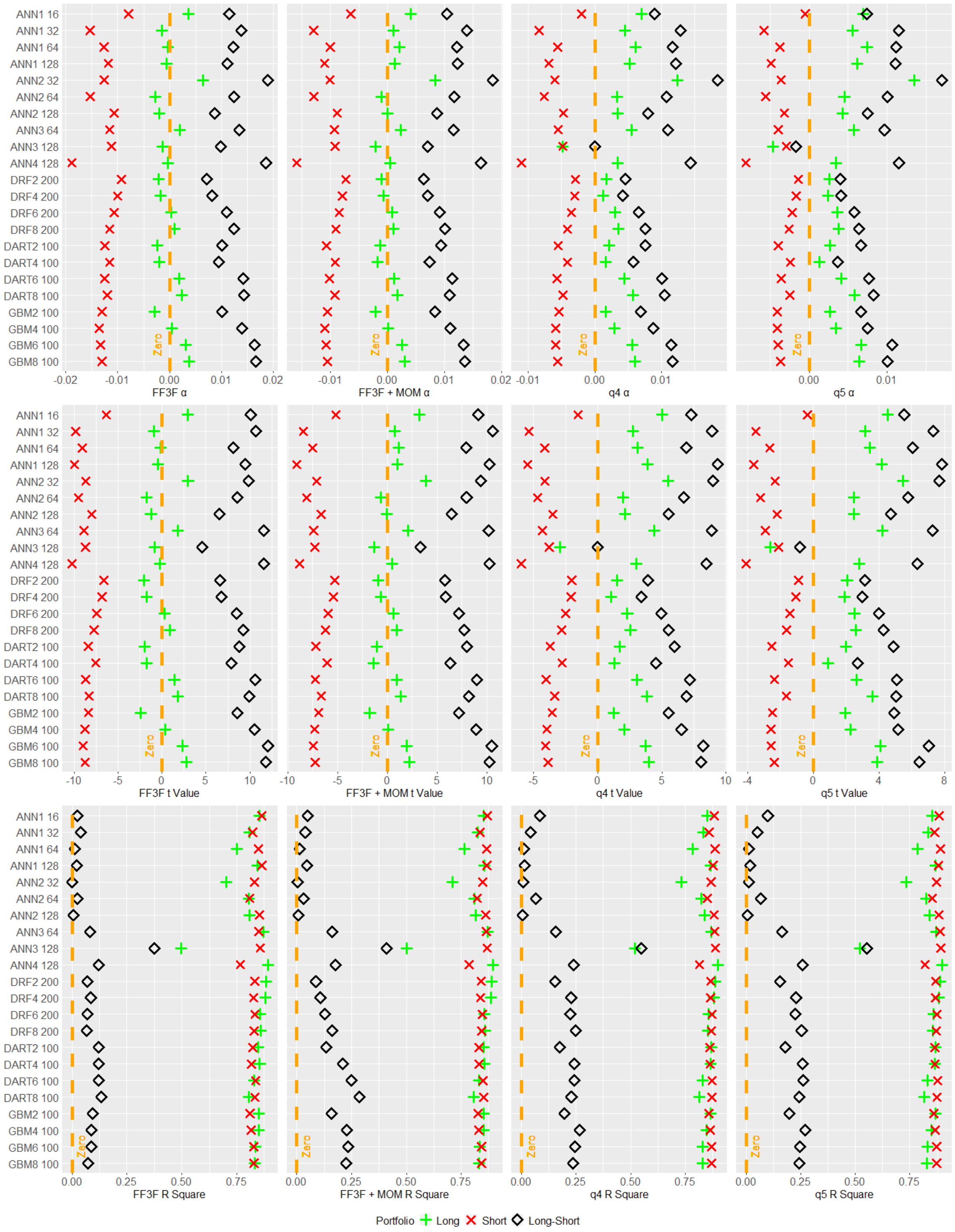

**Figure 9 Factor Model Tests on Value Weight OOS Portfolio Returns 196301:201912**

**Figure 9 (Continues)**

Similar to Figure 8, Figure 9 demonstrates the factor model tests for the value weight portfolios. We also present the model alphas, the t statistics of the model alphas and the R square of the models in the 3 lines of subgraphs. Again, it is obvious that the factor models cannot explain the returns achieved by our long-short portfolios. Taking our long-short portfolio based on our neural network model, ANN2 32, the portfolio has alphas of 1.9 %, 1.85 %, 1.84 %, 1.71 % respectively against FF3F, FF3F + MOM, q4 and q5. The explanatory power of the factor models on the portfolios also seems decreasing as the complexity of our classification models increases.

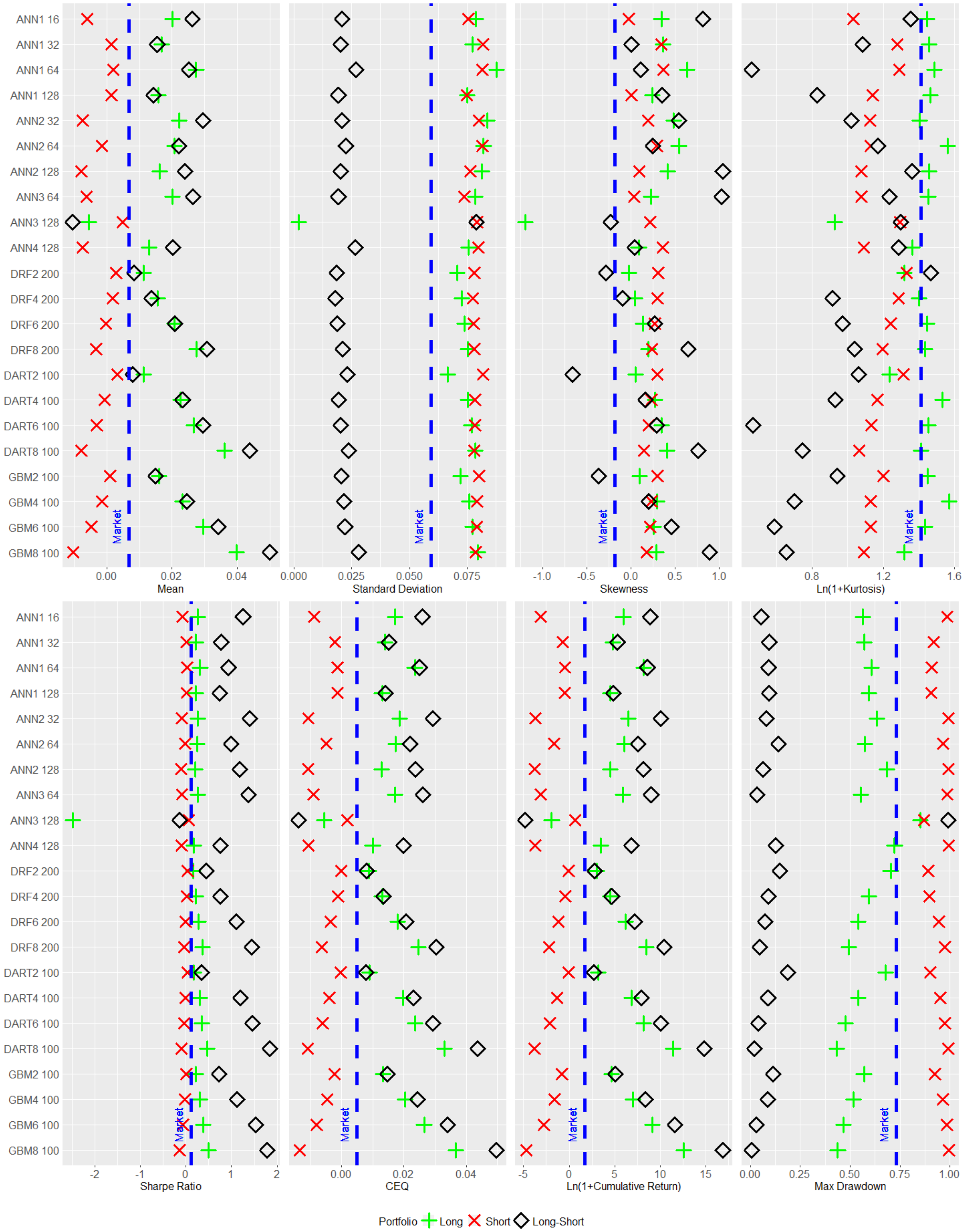

**Figure 10 Time Series IS Equal Weight Portfolio Performance 196301:199112**

**Figure 10 (Continues)**

Figure 10 presents the economic metrics of the in-sample (IS) performance of the equal weight portfolios based on the fitted classification models from 196301:199112. The setup of the figure is the same as in Figure 2. Comparing the OOS performance figures (Figure 2-5) to Figure 10, we show the consistency between the IS performance and the OOS performance. The sample period from 196301:199112 is used to train the model for OOS evaluation covering 199201:201912. The IS performance of the portfolios is similar to the OOS performance but at a magnified level. For example, during the IS period of 196301:199112, the long-short portfolio based on GBM8 100 achieves a monthly SR of 1.7824 while the market delivers a SR of 0.2078. The best IS model in term of risk-return tradeoff in the period from 196301:199112 is the tree model DART8 100, which delivers a monthly SR of 1.8463. Combining Figure 11 and earlier figures on OOS performance, the importance of OOS evaluation is clear.

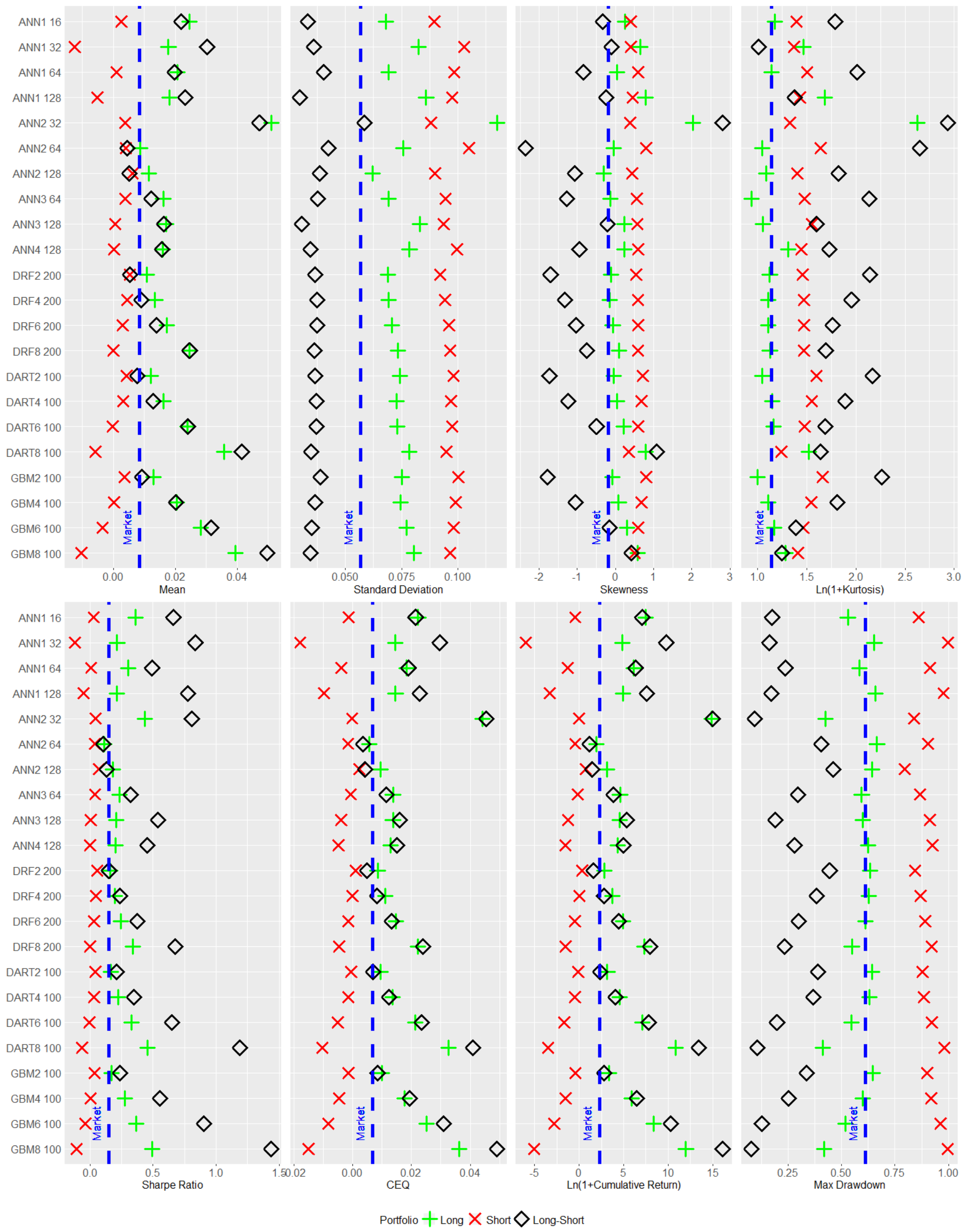

**Figure 11 Time Series IS Equal Weight Portfolio Performance 199201:201912**

**Figure 11 (Continues)**

Figure 11 demonstrates the IS performance of our portfolios in the complemental time period from 199201:201912. The setup of the figure is the same as in Figure 2. The period 199201:201912 is used to train model for OOS evaluation covering 196301:199112. Similar to what is demonstrated in Figure 10, the IS performance during the period from 199201:201912 is at a magnified level.

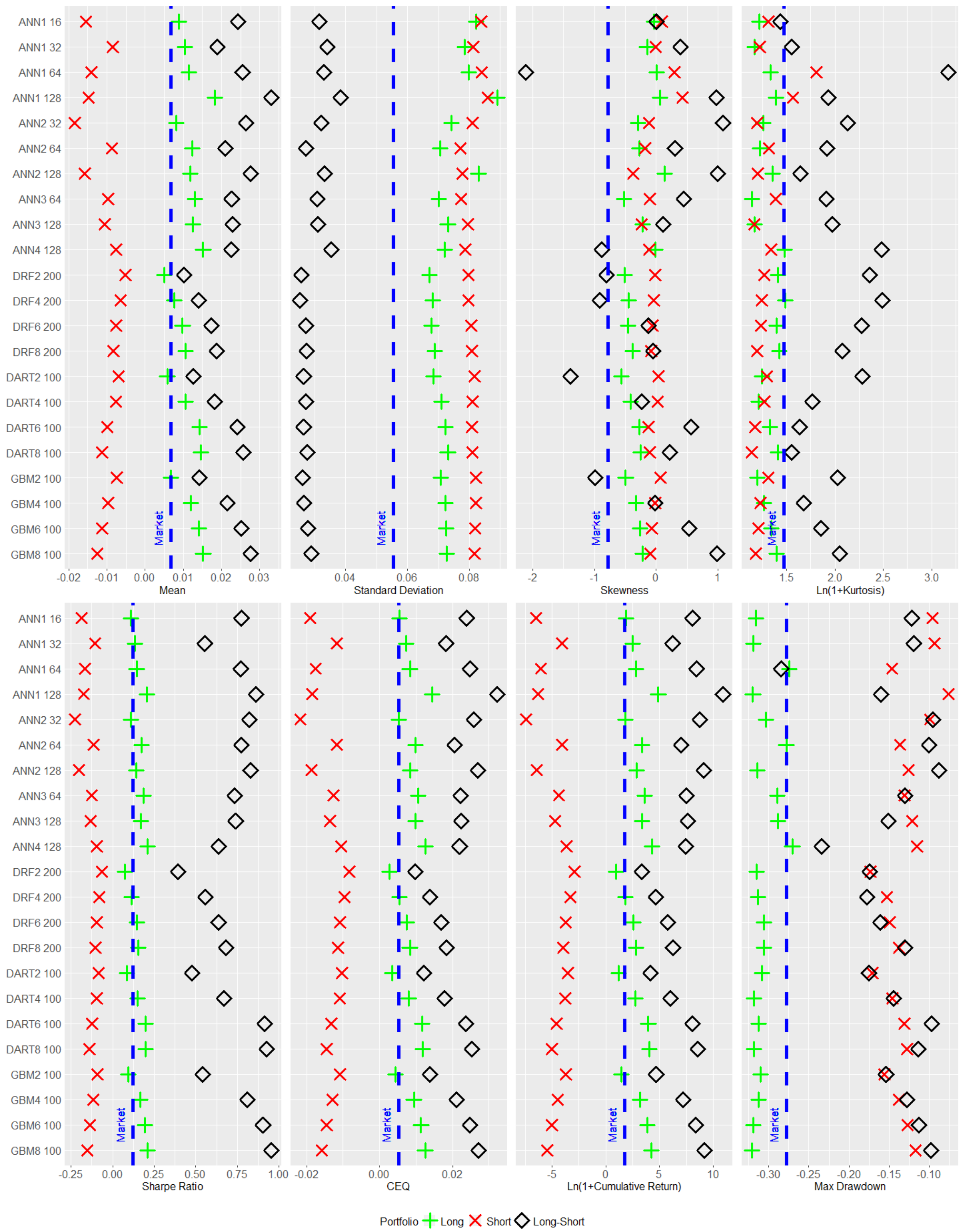

**Figure 12 Cross Sectional OOS Equal Weight Portfolio Performance 196302:201912**

**Figure 12 (Continues)**

Figure 12 presents the economic metrics of the equal weight portfolios performance evaluated based on our cross-sectional OOS setup. Specifically, we split even number months and odd number months during the sample period of 196301:201912. We train our modes based on odd number months and test our models based on even number months. This cross-sectional setup is in the spirit of Fama and French (2018). The cross-sectional OOS evaluation further confirms the superior performance our long-short portfolios based on classification models. The detailed OOS performance of the cross-sectional equal weight portfolios based on the different models mirrors the OOS performance of our equal weight portfolios in the time series test setup as shown in Figure 2.

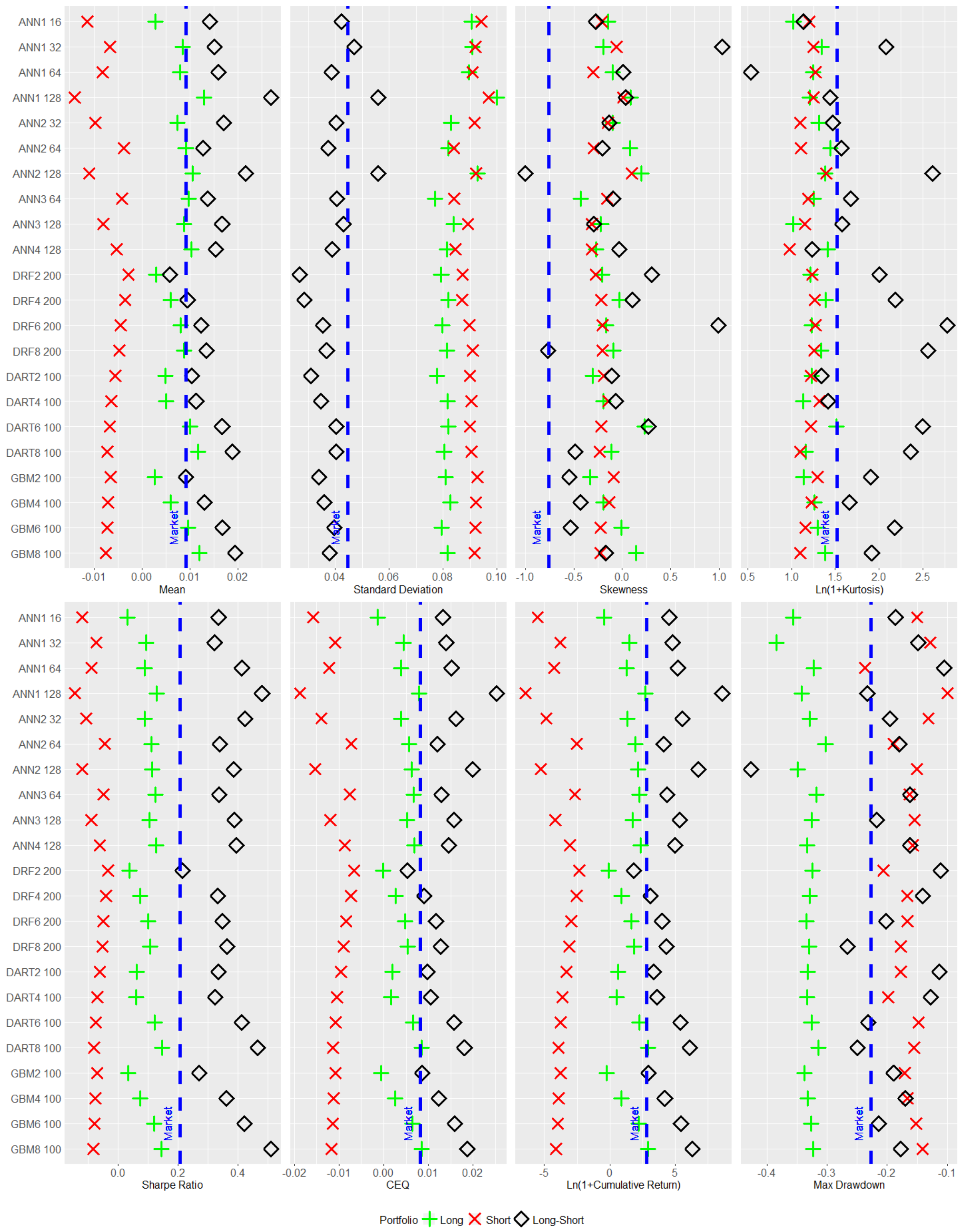

**Figure 13 Cross-sectional OOS Value Weight Portfolio Performance 196302:201912**

**Figure 13 (Continues)**

Figure 13 presents the cross-sectional performance of the portfolios based on our classification models in value weight scheme. Similar to Figure 12, the cross-sectional OOS evaluation confirms the superior performance of our long-short portfolios based on classification models. The detailed OOS performance of the cross-sectional value weight portfolios based on the different models also seems mirroring the OOS performance of our portfolios based on the time series test setup as in Figure 3.

**Table 1 Specification of Return State Classes**

Table 1 describes how we classify return into10 return states. We cross-sectionally rank individual stock returns by trading month, put them into their corresponding deciles and use the deciles as the classes of return states. For example, if a stock falls into the lowest decile in a trading month, we define the true label of the stock as the class of return state 1. A stock in return state 1 means that the stock delivers a return that is among the worst performing returns of the trading month. A stock in return state 10 indicates that the stock are among the stocks delivering the best performing returns of the trading month.

| Specification of Modeling Target | |
|---|---|
| 10 Return States | Criteria |
| 1 | Numeric return less than 10 percentile in a month |
| 2 | Numeric return less than 20 percentile but greater than or equal to 10 percentile in a month |
| 3 | Numeric return less than 30 percentile but greater than or equal to 20 percentile in a month |
| 4 | Numeric return less than 40 percentile but greater than or equal to 30 percentile in a month |
| 5 | Numeric return less than 50 percentile but greater than or equal to 40 percentile in a month |
| 6 | Numeric return less than 60 percentile but greater than or equal to 50 percentile in a month |
| 7 | Numeric return less than 70 percentile but greater than or equal to 60 percentile in a month |
| 8 | Numeric return less than 80 percentile but greater than or equal to 70 percentile in a month |
| 9 | Numeric return less than 90 percentile but greater than or equal to 80 percentile in a month |
| 10 | Numeric return greater than or equal to 90 percentile in a month |

**Table 2 Model Specification**

Table 2 presents the description of models we apply in this study. Overall, we include 2 modeling architectures and 22 models. Panel A demonstrates the architectural specification for each model. Panel B demonstrates the additional specifications for our neural network models. Panel C demonstrates the hyperparameters that we choose based on cross-validation. Section 2.1.3 discusses the details for each model. Note that we use enum encoding for the categorical variables in our tree models. One-hot-encoding splits the categorical variables into smaller binary choice questions and marks the associated values as positive indicated by 1 and negative indicated by 0. The enum encoding considers the categorical values as nonordinal values. In the tree models, different categories will have different leaves. We do not use one-hot-encoding as it is not the native choice of the tree models and will introduce sparsity which lowers the information ratio for the tree models.

Panel A: Architectural Specifications

| | | | Structural | Structural |
|---|---|---|---|---|
| Model | Architecture | Specification | Complexity | Capacity |
| ANN1 128 | Neuron Network | Multilayer Perceptron | 1 Hidden Layer | # Neurons = 128 |
| ANN1 16 | Neuron Network | Multilayer Perceptron | 1 Hidden Layer | # Neurons = 16 |
| ANN1 32 | Neuron Network | Multilayer Perceptron | 1 Hidden Layer | # Neurons = 32 |
| ANN1 64 | Neuron Network | Multilayer Perceptron | 1 Hidden Layer | # Neurons = 64 |
| ANN2 128 | Neuron Network | Multilayer Perceptron | 2 Hidden Layers | # Neurons = {128,64} |
| ANN2 32 | Neuron Network | Multilayer Perceptron | 2 Hidden Layers | # Neurons = {32,16} |
| ANN2 64 | Neuron Network | Multilayer Perceptron | 2 Hidden Layers | # Neurons = {64,32} |
| ANN3 128 | Neuron Network | Multilayer Perceptron | 3 Hidden Layers | # Neurons = {128,64,32} |
| ANN3 64 | Neuron Network | Multilayer Perceptron | 3 Hidden Layers | # Neurons = {64,32,16} |
| ANN4 128 | Neuron Network | Multilayer Perceptron | 4 Hidden Layers | # Neurons = {128,64,32,16} |
| DART2 100 | Tree | Boosting Tree | Maximum Depth = 2 | # Trees = 100 |
| DART4 100 | Tree | Boosting Tree | Maximum Depth = 4 | # Trees = 100 |
| DART6 100 | Tree | Boosting Tree | Maximum Depth = 6 | # Trees = 100 |
| DART8 100 | Tree | Boosting Tree | Maximum Depth = 8 | # Trees = 100 |
| DRF2 200 | Tree | Forest | Maximum Depth = 2 | # Trees = 200 |
| DRF4 200 | Tree | Forest | Maximum Depth = 4 | # Trees = 200 |
| DRF6 200 | Tree | Forest | Maximum Depth = 6 | # Trees = 200 |
| DRF8 200 | Tree | Forest | Maximum Depth = 8 | # Trees = 200 |
| GBM2 100 | Tree | Boosting Tree | Maximum Depth = 2 | # Trees = 100 |
| GBM4 100 | Tree | Boosting Tree | Maximum Depth = 4 | # Trees = 100 |
| GBM6 100 | Tree | Boosting Tree | Maximum Depth = 6 | # Trees = 100 |
| GBM8 100 | Tree | Boosting Tree | Maximum Depth = 8 | # Trees = 100 |

Panel B: ANN Other Specifications

| Hidden | Output | Categorical | | |
|---|---|---|---|---|
| Layer | Layer | Variable | # epochs | Loss |
| Activation | Activation | Encoding | | |
| Tanh | Softmax | One hot encoding | 50 | Cross Entropy |

Panel C: Hyperparameters

| Architecture | Model | Parameter | Candidate |
|---|---|---|---|
| Neuron Network | All | L1 Regulation | 0.01, 0.001, 0.0001, 0.00001 |
| Tree | All | Sample Rate | 0.8. 1 |
| Tree | All | Predictor Sample Rate | 0.8, 1 |

**Table 3 Selection of No Information Benchmark Classifier**

Table 3 presents the Tukey's HSD multiple comparison test with Monte Carlo simulation. The testing samples are generated with the sample covering 199201:201912. Classifier 1 is the random classifier that assigns return states with equal probability. Classifier 2 is the random classifier that assigns return states with IS sample probability mass function observed in the sample 196301:199112. Classifier 3 is the naïve classifier that assigns the most populated IS return state to all OOS observations. Classifier 4 is the random classifier that assigns return states with OOS sample probability mass function with equal probability. Classifier 5 is the naïve classifier that assigns the most populated OOS returns state to all OOS observations. Note that even with minimum information, the naïve classifier which assigns the most populated IS return state to all OOS observations demonstrates higher accuracy than random classifier that uses no information. To provide a comprehensive evaluation of the proper benchmarks, we also consider the martingale hypothesis about return process, i.e., the best prediction for the future return is today's return and the return process is a memoryless process. We produce Classifier 6 to account for the martingale hypothesis by predicting the future return state with the current return state. Because of introducing historical return state information, Classifier 6 has better overall accuracy in our simulation. We include both Classifier 5 and Classifier 6 as our benchmarks in our binomial tests.

| | Difference | Lower 95% Bound | Upper 95% Bound | P Value |
|---|---|---|---|---|
| 1-2 | 0.0000 | -0.0002 | 0.0002 | 1.0000 |
| 1-3 | 0.0004 | 0.0002 | 0.0006 | 0.0000 |
| 1-4 | 0.0000 | -0.0001 | 0.0002 | 0.9671 |
| 1-5 | 0.0005 | 0.0004 | 0.0007 | 0.0000 |
| 1-6 | 0.0201 | 0.0199 | 0.0203 | 0.0000 |
| 2-3 | 0.0004 | 0.0002 | 0.0005 | 0.0000 |
| 2-4 | 0.0000 | -0.0001 | 0.0002 | 0.9836 |
| 2-5 | 0.0005 | 0.0004 | 0.0007 | 0.0000 |
| 2-6 | 0.0201 | 0.0199 | 0.0203 | 0.0000 |
| 3-4 | -0.0003 | -0.0005 | -0.0002 | 0.0000 |
| 3-5 | 0.0002 | 0.0000 | 0.0003 | 0.1138 |
| 3-6 | 0.0197 | 0.0196 | 0.0199 | 0.0000 |
| 4-5 | 0.0005 | 0.0003 | 0.0007 | 0.0000 |
| 4-6 | 0.0201 | 0.0199 | 0.0202 | 0.0000 |
| 5-6 | 0.0196 | 0.0194 | 0.0197 | 0.0000 |

**Table 4 Data Construction Summary Statistics**

Table 4 presents the summary statistics of our data with CRSP database as the reference. Panel A presents number of securities in our sample. Panels B and C present summary statistics and market capitalization in month t-1, respectively.

| Panel A: Number of Securities Summary | | | | | |
|---|---|---|---|---|---|
| Sample | Distinct Total | Mean | Min | Max | Filter |
| CRSP | 33004 | 6146.905 | 2069 | 9366 | None |
| Our Sample | 26302 | 4886.6754 | 1997 | 7929 | No missing return; EXCHCD and SHRCD |

| Panel B: Summary Statistics of Returns | | | | | |
|---|---|---|---|---|---|
| Sample | Mean | SD | Skewness | Kurtosis | Filter |
| CRSP | 0.0102 | 0.176 | 20.8963 | 5165.1519 | No missing return |
| Our Sample | 0.0109 | 0.1883 | 20.6107 | 4785.8618 | No missing return; EXCHCD and SHRCD |

| Panel C: Summary Statistics of Market Capitalization at t-1 | | | | | |
|---|---|---|---|---|---|
| Sample | Mean | SD | Skewness | Kurtosis | Filter |
| CRSP | 1601233.289 | 11003218.89 | 27.6035 | 1369.1445 | No missing t-1 ME |
| Our Sample | 1723086.927 | 11923239.69 | 26.0793 | 1202.901 | No missing t-1 ME; EXCHCD and SHRCD |

**Table 5 Sample Splitting**

Table 5 describes the period we apply for training and testing setup for in-sample (IS) and out-of-sample (OOS). Specifically, we form an overall time series OOS test sample covering 196301:201912 based on splitting the time period into two time series training periods. We use the time period from 196301:199112 to train models for OOS predictions in the period from 199201:201912 and we use the time period from 199201:201912 to train models for OOS predictions in the period from 196301:199112. We combine the OOS predictions for OOS evaluation. In the spirit of Fama and French (2018), we also split the data by even number and odd number months for the cross-sectional OOS evaluation. We train our models cross-sectionally with odd number months from 196301:201911 and test the OOS predictions with the even number months from 196302:201912. Our data splitting is in the spirit of Fama and French (2018) and Martin and Nagel (2020).

| Training and Testing Setup | IS Training | OOS Testing |
|---|---:|---:|
| Time Series Setup 1 | 196301:199112 | 199201:201912 |
| Time Series Setup 2 | 199201:201912 | 196301:199112 |
| Time Series Main Setup<br>(Combined Cross-Validation) | 196301:199112 and 199201:201912 | 199201:201912 and 196301:199112 |
| Cross-Sectional Setup | Odd Number Months 196301:201911 | Even Number Months 196302:201912 |

**Table 6 Comparison on Economic Performance between 2 Data Setups 196301:201912**

We demonstrate the economic performance comparison across our 2 data setups, i.e., portfolios based on data sets with and without macro data components. For results in Panel A and B, we include numeric characteristics based long-short factor mimicking returns, 9 predictors from the data set of Goyal and Welch (2009) and 125 macroeconomic indicators from FRED-MD. Our results in Panel C and D do not include these macro-level data components. The performance of results does not change much for ANN models, despite that the best models do change as different dimensionality does impact ANNs' performance heavily. However, the performance for tree models increases with less macro-level predictors, indicating that the contribution for individual stock return state prediction is mainly from characteristics. Because of performance similarity, we present our latter tables for time series OOS setup with data excluding macro level data components.

Panel A: Equal Weight Long-Short Portfolio Economic Performance with Macro Components

| Model | Mean | SD | Skewness | SR | CEQ | Cumulative Return | Min | Max DD |
|---|---|---|---|---|---|---|---|---|
| Buy-Hold | 0.01 | 0.06 | -0.19 | 0.13 | 0.01 | 5478% | -0.28 | 0.73 |
| ANN1 128 | 0.02 | 0.03 | 1.71 | 0.82 | 0.02 | 726187870% | -0.05 | 0.06 |
| ANN1 16 | 0.02 | 0.03 | 1.39 | 0.62 | 0.02 | 3684032% | -0.06 | 0.17 |
| ANN1 32 | 0.02 | 0.02 | 0.09 | 0.65 | 0.01 | 2161951% | -0.08 | 0.15 |
| ANN1 64 | 0.03 | 0.03 | 1.82 | 0.88 | 0.03 | 3666944052% | -0.07 | 0.08 |
| ANN2 128 | 0.01 | 0.04 | 0.00 | 0.33 | 0.01 | 274550% | -0.12 | 0.46 |
| ANN2 32 | 0.02 | 0.03 | -0.42 | 0.53 | 0.02 | 4518415% | -0.16 | 0.25 |
| ANN2 64 | 0.02 | 0.03 | 0.60 | 0.74 | 0.02 | 38292152% | -0.07 | 0.08 |
| ANN3 128 | 0.01 | 0.06 | -0.82 | 0.09 | 0.00 | 1047% | -0.40 | 0.99 |
| ANN3 64 | 0.02 | 0.03 | -0.46 | 0.66 | 0.02 | 7567516% | -0.12 | 0.18 |
| ANN4 128 | 0.01 | 0.03 | -0.42 | 0.40 | 0.01 | 268570% | -0.13 | 0.29 |
| DART2 100 | 0.01 | 0.02 | -0.17 | 0.24 | 0.01 | 3134% | -0.10 | 0.19 |
| DART4 100 | 0.01 | 0.03 | 1.25 | 0.36 | 0.01 | 88745% | -0.16 | 0.33 |
| DART6 100 | 0.01 | 0.03 | 0.99 | 0.44 | 0.01 | 731396% | -0.19 | 0.37 |
| DART8 100 | 0.02 | 0.03 | 0.07 | 0.60 | 0.02 | 6790642% | -0.13 | 0.31 |
| DRF2 200 | 0.01 | 0.03 | -1.52 | 0.18 | 0.00 | 2325% | -0.22 | 0.29 |
| DRF4 200 | 0.01 | 0.03 | -1.17 | 0.32 | 0.01 | 27969% | -0.16 | 0.30 |
| DRF6 200 | 0.01 | 0.03 | -0.76 | 0.43 | 0.01 | 219281% | -0.16 | 0.31 |
| DRF8 200 | 0.01 | 0.03 | -0.72 | 0.50 | 0.01 | 985394% | -0.14 | 0.32 |
| GBM2 100 | 0.01 | 0.03 | -0.87 | 0.32 | 0.01 | 21317% | -0.16 | 0.30 |
| GBM4 100 | 0.02 | 0.03 | -0.40 | 0.55 | 0.01 | 2209709% | -0.16 | 0.28 |
| GBM6 100 | 0.02 | 0.03 | -0.03 | 0.69 | 0.02 | 46490610% | -0.14 | 0.22 |
| GBM8 100 | 0.02 | 0.03 | -0.02 | 0.76 | 0.02 | 132529676% | -0.14 | 0.19 |

Panel B: Value Weight Long-Short Portfolio Economic Performance with Macro Components

| Model | Mean | SD | Skewness | SR | CEQ | Cumulative Return | Min | Max DD |
|---|---|---|---|---|---|---|---|---|
| Buy-Hold | 0.01 | 0.04 | -0.56 | 0.12 | 0.00 | 2004% | -0.23 | 0.55 |
| ANN1 128 | 0.01 | 0.04 | 0.23 | 0.40 | 0.01 | 1378238% | -0.15 | 0.32 |
| ANN1 16 | 0.01 | 0.03 | 0.20 | 0.42 | 0.01 | 822463% | -0.14 | 0.28 |
| ANN1 32 | 0.01 | 0.03 | -0.04 | 0.40 | 0.01 | 174110% | -0.11 | 0.24 |
| ANN1 64 | 0.02 | 0.04 | 0.06 | 0.45 | 0.02 | 3648840% | -0.15 | 0.23 |
| ANN2 128 | 0.01 | 0.06 | 0.30 | 0.26 | 0.01 | 693118% | -0.19 | 0.65 |
| ANN2 32 | 0.02 | 0.04 | -0.69 | 0.39 | 0.02 | 6118266% | -0.23 | 0.39 |
| ANN2 64 | 0.02 | 0.04 | -0.51 | 0.40 | 0.02 | 4549616% | -0.23 | 0.50 |
| ANN3 128 | 0.01 | 0.06 | 0.08 | 0.08 | 0.00 | 751% | -0.28 | 0.92 |
| ANN3 64 | 0.02 | 0.04 | -0.28 | 0.43 | 0.02 | 3402828% | -0.17 | 0.38 |
| ANN4 128 | 0.01 | 0.04 | -0.29 | 0.37 | 0.01 | 1340309% | -0.16 | 0.34 |
| DART2 100 | 0.01 | 0.03 | 0.59 | 0.21 | 0.01 | 5650% | -0.11 | 0.34 |
| DART4 100 | 0.01 | 0.04 | -0.37 | 0.18 | 0.01 | 6197% | -0.18 | 0.67 |
| DART6 100 | 0.01 | 0.05 | -0.55 | 0.23 | 0.01 | 63101% | -0.23 | 0.66 |
| DART8 100 | 0.01 | 0.03 | -0.35 | 0.31 | 0.01 | 115740% | -0.12 | 0.42 |
| DRF2 200 | 0.00 | 0.03 | 0.23 | 0.16 | 0.00 | 1548% | -0.13 | 0.27 |
| DRF4 200 | 0.01 | 0.03 | -0.53 | 0.20 | 0.01 | 5555% | -0.19 | 0.34 |
| DRF6 200 | 0.01 | 0.03 | -0.37 | 0.30 | 0.01 | 61308% | -0.16 | 0.34 |
| DRF8 200 | 0.01 | 0.04 | -0.94 | 0.34 | 0.01 | 223733% | -0.22 | 0.36 |
| GBM2 100 | 0.01 | 0.03 | -0.52 | 0.28 | 0.01 | 35243% | -0.17 | 0.43 |
| GBM4 100 | 0.01 | 0.04 | -0.46 | 0.37 | 0.01 | 534859% | -0.17 | 0.51 |
| GBM6 100 | 0.02 | 0.04 | -0.51 | 0.40 | 0.02 | 2561820% | -0.17 | 0.47 |
| GBM8 100 | 0.02 | 0.04 | -0.56 | 0.40 | 0.02 | 2573228% | -0.19 | 0.41 |

**Table 6 (Continues)**

Panel C: Equal Weight Long-Short Portfolio Economic Performance without Macro Components

| Model | Mean | SD | Skewness | SR | CEQ | Cumulative Return | Min | Max DD |
|---|---|---|---|---|---|---|---|---|
| Buy-Hold | 0.01 | 0.06 | -0.19 | 0.13 | 0.01 | 5478% | -0.28 | 0.73 |
| ANN1 128 | 0.02 | 0.02 | -0.28 | 0.63 | 0.02 | 3280169% | -0.14 | 0.28 |
| ANN1 16 | 0.02 | 0.02 | 0.52 | 0.82 | 0.02 | 15575952% | -0.10 | 0.13 |
| ANN1 32 | 0.02 | 0.03 | -0.70 | 0.61 | 0.02 | 9033903% | -0.18 | 0.19 |
| ANN1 64 | 0.02 | 0.03 | 0.51 | 0.60 | 0.02 | 4309857% | -0.08 | 0.25 |
| ANN2 128 | 0.01 | 0.02 | 0.12 | 0.50 | 0.01 | 295798% | -0.09 | 0.15 |
| ANN2 32 | 0.03 | 0.04 | 1.29 | 0.87 | 0.03 | 485649224179% | -0.07 | 0.14 |
| ANN2 64 | 0.01 | 0.03 | -0.27 | 0.35 | 0.01 | 92522% | -0.14 | 0.39 |
| ANN3 128 | 0.00 | 0.06 | -1.06 | 0.07 | 0.00 | 384% | -0.47 | 0.99 |
| ANN3 64 | 0.02 | 0.02 | 0.26 | 0.72 | 0.02 | 4897082% | -0.09 | 0.11 |
| ANN4 128 | 0.02 | 0.03 | -0.18 | 0.64 | 0.02 | 28746938% | -0.11 | 0.21 |
| DART2 100 | 0.01 | 0.03 | -1.27 | 0.25 | 0.01 | 7284% | -0.16 | 0.34 |
| DART4 100 | 0.01 | 0.03 | -0.38 | 0.40 | 0.01 | 123984% | -0.17 | 0.31 |
| DART6 100 | 0.02 | 0.03 | 0.22 | 0.60 | 0.02 | 6861226% | -0.14 | 0.22 |
| DART8 100 | 0.02 | 0.03 | 0.42 | 0.72 | 0.02 | 50429123% | -0.10 | 0.15 |
| DRF2 200 | 0.01 | 0.03 | -1.56 | 0.19 | 0.00 | 2417% | -0.20 | 0.30 |
| DRF4 200 | 0.01 | 0.03 | -1.00 | 0.31 | 0.01 | 29705% | -0.16 | 0.33 |
| DRF6 200 | 0.01 | 0.03 | -0.65 | 0.46 | 0.01 | 458662% | -0.18 | 0.34 |
| DRF8 200 | 0.02 | 0.03 | -0.39 | 0.59 | 0.02 | 4727097% | -0.17 | 0.31 |
| GBM2 100 | 0.01 | 0.03 | -0.82 | 0.33 | 0.01 | 24558% | -0.16 | 0.29 |
| GBM4 100 | 0.02 | 0.03 | -0.27 | 0.56 | 0.02 | 3028396% | -0.15 | 0.28 |
| GBM6 100 | 0.02 | 0.03 | 0.17 | 0.73 | 0.02 | 64231407% | -0.12 | 0.18 |
| GBM8 100 | 0.02 | 0.03 | 0.46 | 0.81 | 0.02 | 200202148% | -0.07 | 0.12 |

Panel D: Value Weight Long-Short Portfolio Economic Performance without Macro Components

| Model | Mean | SD | Skewness | SR | CEQ | Cumulative Return | Min | Max DD |
|---|---|---|---|---|---|---|---|---|
| Buy-Hold | 0.01 | 0.04 | -0.56 | 0.12 | 0.00 | 2004% | -0.23 | 0.55 |
| ANN1 128 | 0.01 | 0.03 | 0.39 | 0.37 | 0.01 | 172108% | -0.10 | 0.25 |
| ANN1 16 | 0.01 | 0.03 | 0.24 | 0.39 | 0.01 | 183825% | -0.11 | 0.20 |
| ANN1 32 | 0.01 | 0.03 | 0.13 | 0.42 | 0.01 | 1103697% | -0.13 | 0.29 |
| ANN1 64 | 0.01 | 0.04 | -0.10 | 0.33 | 0.01 | 352299% | -0.19 | 0.49 |
| ANN2 128 | 0.01 | 0.03 | 0.53 | 0.25 | 0.01 | 22764% | -0.10 | 0.26 |
| ANN2 32 | 0.02 | 0.05 | 0.47 | 0.39 | 0.02 | 20738716% | -0.17 | 0.30 |
| ANN2 64 | 0.01 | 0.04 | 0.14 | 0.32 | 0.01 | 220580% | -0.15 | 0.43 |
| ANN3 128 | 0.01 | 0.07 | -0.08 | 0.08 | 0.00 | 821% | -0.29 | 0.96 |
| ANN3 64 | 0.01 | 0.03 | -0.13 | 0.39 | 0.01 | 293578% | -0.12 | 0.32 |
| ANN4 128 | 0.02 | 0.04 | -0.70 | 0.38 | 0.02 | 4784931% | -0.30 | 0.40 |
| DART2 100 | 0.01 | 0.03 | -0.09 | 0.28 | 0.01 | 31399% | -0.13 | 0.27 |
| DART4 100 | 0.01 | 0.03 | -0.20 | 0.26 | 0.01 | 21838% | -0.15 | 0.44 |
| DART6 100 | 0.01 | 0.04 | -0.45 | 0.35 | 0.01 | 429355% | -0.22 | 0.31 |
| DART8 100 | 0.01 | 0.04 | -0.63 | 0.33 | 0.01 | 415020% | -0.23 | 0.44 |
| DRF2 200 | 0.01 | 0.03 | 0.06 | 0.21 | 0.01 | 4522% | -0.15 | 0.29 |
| DRF4 200 | 0.01 | 0.03 | -0.24 | 0.22 | 0.01 | 9153% | -0.15 | 0.39 |
| DRF6 200 | 0.01 | 0.03 | -0.83 | 0.29 | 0.01 | 62025% | -0.22 | 0.39 |
| DRF8 200 | 0.01 | 0.04 | -1.12 | 0.32 | 0.01 | 157852% | -0.25 | 0.51 |
| GBM2 100 | 0.01 | 0.03 | -0.41 | 0.29 | 0.01 | 34676% | -0.15 | 0.39 |
| GBM4 100 | 0.01 | 0.04 | -0.57 | 0.36 | 0.01 | 392902% | -0.19 | 0.58 |
| GBM6 100 | 0.02 | 0.04 | -0.31 | 0.42 | 0.01 | 1988854% | -0.15 | 0.46 |
| GBM8 100 | 0.02 | 0.04 | 0.05 | 0.42 | 0.02 | 2653376% | -0.15 | 0.47 |

**Table 7 Overall Accuracy of Time Series OOS Prediction and Binomial Test 196301:201912**

Table 7 presents the accuracy of each model. The accuracy of a model is the direct evaluation of the correctness of the model predictions. The Kappa statistic measures the level of agreement between the predictions and the actual data and higher Kappa statistic indicates better performance. According to Landis and Koch (1977), a Kappa value greater than 0 but less than 0.2 indicates that the agreement is slight. The confidence interval is the binomial confidence interval based on accuracy. The P values are associated with the hypothesis test on whether the accuracy is different from the 2 benchmark accuracies statistically. We discussed our model specifications and the statistical metrics in Section 2. Table 7 shows that all of our models are better than the no information accuracy which is calculated under the assumption that the historical information is useless in terms of prediction future return states. All of our models are also better than the martingale accuracy, which is calculated under the assumption that the stock returns follow a memoryless process.

| Model | Accuracy | Kappa | Lower 99% Bound | Upper 99% Bound | No Info Accuracy | No Info P Value | Martingale Accuracy | Martingale P Value |
|---|---|---|---|---|---|---|---|---|
| ANN1 16 | 0.153 | 0.059 | 0.153 | 0.154 | 0.102 | **0.000** | 0.117 | **0.000** |
| ANN1 32 | 0.153 | 0.058 | 0.153 | 0.154 | 0.102 | **0.000** | 0.117 | **0.000** |
| ANN1 64 | 0.154 | 0.059 | 0.153 | 0.154 | 0.102 | **0.000** | 0.117 | **0.000** |
| ANN1 128 | 0.150 | 0.056 | 0.150 | 0.151 | 0.102 | **0.000** | 0.117 | **0.000** |
| ANN2 32 | 0.153 | 0.058 | 0.153 | 0.154 | 0.102 | **0.000** | 0.117 | **0.000** |
| ANN2 64 | 0.151 | 0.056 | 0.151 | 0.152 | 0.102 | **0.000** | 0.117 | **0.000** |
| ANN2 128 | 0.152 | 0.057 | 0.151 | 0.152 | 0.102 | **0.000** | 0.117 | **0.000** |
| ANN3 64 | 0.152 | 0.058 | 0.152 | 0.153 | 0.102 | **0.000** | 0.117 | **0.000** |
| ANN3 128 | 0.151 | 0.056 | 0.150 | 0.151 | 0.102 | **0.000** | 0.117 | **0.000** |
| ANN4 128 | 0.152 | 0.058 | 0.152 | 0.153 | 0.102 | **0.000** | 0.117 | **0.000** |
| DART2 100 | 0.153 | 0.059 | 0.153 | 0.154 | 0.102 | **0.000** | 0.117 | **0.000** |
| DART4 100 | 0.155 | 0.061 | 0.155 | 0.156 | 0.102 | **0.000** | 0.117 | **0.000** |
| DART6 100 | 0.156 | 0.062 | 0.156 | 0.157 | 0.102 | **0.000** | 0.117 | **0.000** |
| DART8 100 | 0.156 | 0.062 | 0.155 | 0.156 | 0.102 | **0.000** | 0.117 | **0.000** |
| DRF2 200 | 0.152 | 0.057 | 0.152 | 0.153 | 0.102 | **0.000** | 0.117 | **0.000** |
| DRF4 200 | 0.156 | 0.061 | 0.155 | 0.156 | 0.102 | **0.000** | 0.117 | **0.000** |
| DRF6 200 | 0.157 | 0.063 | 0.157 | 0.158 | 0.102 | **0.000** | 0.117 | **0.000** |
| DRF8 200 | 0.158 | 0.064 | 0.158 | 0.159 | 0.102 | **0.000** | 0.117 | **0.000** |
| GBM2 100 | 0.155 | 0.061 | 0.155 | 0.156 | 0.102 | **0.000** | 0.117 | **0.000** |
| GBM4 100 | 0.157 | 0.063 | 0.157 | 0.158 | 0.102 | **0.000** | 0.117 | **0.000** |
| GBM6 100 | 0.158 | 0.064 | 0.157 | 0.159 | 0.102 | **0.000** | 0.117 | **0.000** |
| GBM8 100 | 0.158 | 0.064 | 0.157 | 0.158 | 0.102 | **0.000** | 0.117 | **0.000** |

**Table 8 Return State Transition Probability and Mean Return 196301:201912**

This table presents the return state transition probability and mean return of the transition from the old to new state. A stock in return state 1 means that the stock delivers a return that is among the worst performing returns of the trading month. A stock in return state 10 indicates that the stock are among the stocks delivering the best performing returns of the trading month. Note that the true return state transition probabilities are not evenly distributed. The extreme return states and the middle return states are associated with transition probabilities either substantially greater than 10 % or substantially lower than 10%. These states are thus with higher certainty in the process of state transition. More specifically, return state 3, return state 4 and return state 9 seem the most uncertain states. The return state 1 and return state 10 seem the most certain states. The return states are defined in Table 1.

Panel A: True Return State Transition Probability Matrix 196301:201912

| | New 1 | New 2 | New 3 | New 4 | New 5 | New 6 | New 7 | New 8 | New 9 | New 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Old 1 | 0.1741 | 0.1063 | 0.0816 | 0.0686 | 0.0665 | 0.0660 | 0.0719 | 0.0817 | 0.1052 | 0.1782 |
| Old 2 | 0.1137 | 0.1073 | 0.0963 | 0.0891 | 0.0879 | 0.0875 | 0.0918 | 0.0993 | 0.1090 | 0.1180 |
| Old 3 | 0.0859 | 0.0987 | 0.0997 | 0.1011 | 0.1007 | 0.1033 | 0.1050 | 0.1054 | 0.1059 | 0.0944 |
| Old 4 | 0.0713 | 0.0899 | 0.1007 | 0.1073 | 0.1127 | 0.1134 | 0.1122 | 0.1092 | 0.1014 | 0.0817 |
| Old 5 | 0.0696 | 0.0860 | 0.0992 | 0.1094 | 0.1128 | 0.1203 | 0.1167 | 0.1098 | 0.0981 | 0.0779 |
| Old 6 | 0.0690 | 0.0868 | 0.1002 | 0.1084 | 0.1138 | 0.1177 | 0.1186 | 0.1116 | 0.0970 | 0.0768 |
| Old 7 | 0.0675 | 0.0897 | 0.1025 | 0.1083 | 0.1134 | 0.1163 | 0.1164 | 0.1121 | 0.0980 | 0.0758 |
| Old 8 | 0.0753 | 0.0973 | 0.1054 | 0.1067 | 0.1102 | 0.1123 | 0.1092 | 0.1058 | 0.0984 | 0.0794 |
| Old 9 | 0.0958 | 0.1103 | 0.1061 | 0.1023 | 0.0976 | 0.0974 | 0.0971 | 0.1009 | 0.0999 | 0.0927 |
| Old 10 | 0.1742 | 0.1236 | 0.0966 | 0.0825 | 0.0752 | 0.0736 | 0.0743 | 0.0802 | 0.0912 | 0.1284 |

Panel B: Return State Transition Mean Return 196301:201912

| | New 1 | New 2 | New 3 | New 4 | New 5 | New 6 | New 7 | New 8 | New 9 | New 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Old 1 | -0.2744 | -0.1217 | -0.0750 | -0.0413 | -0.0130 | 0.0114 | 0.0394 | 0.0759 | 0.1350 | 0.4003 |
| Old 2 | -0.2424 | -0.1177 | -0.0716 | -0.0394 | -0.0127 | 0.0127 | 0.0404 | 0.0751 | 0.1292 | 0.3169 |
| Old 3 | -0.2316 | -0.1139 | -0.0691 | -0.0370 | -0.0121 | 0.0126 | 0.0388 | 0.0719 | 0.1243 | 0.2961 |
| Old 4 | -0.2254 | -0.1105 | -0.0661 | -0.0357 | -0.0108 | 0.0119 | 0.0375 | 0.0683 | 0.1193 | 0.2871 |
| Old 5 | -0.2229 | -0.1078 | -0.0624 | -0.0332 | -0.0099 | 0.0120 | 0.0357 | 0.0663 | 0.1174 | 0.2940 |
| Old 6 | -0.2185 | -0.1055 | -0.0614 | -0.0328 | -0.0094 | 0.0127 | 0.0358 | 0.0660 | 0.1165 | 0.2959 |
| Old 7 | -0.2123 | -0.1041 | -0.0623 | -0.0340 | -0.0103 | 0.0114 | 0.0348 | 0.0646 | 0.1128 | 0.2830 |
| Old 8 | -0.2097 | -0.1048 | -0.0627 | -0.0350 | -0.0110 | 0.0113 | 0.0370 | 0.0668 | 0.1168 | 0.2884 |
| Old 9 | -0.2120 | -0.1076 | -0.0664 | -0.0374 | -0.0125 | 0.0113 | 0.0375 | 0.0690 | 0.1207 | 0.2983 |
| Old 10 | -0.2321 | -0.1137 | -0.0707 | -0.0406 | -0.0135 | 0.0115 | 0.0378 | 0.0714 | 0.1276 | 0.3506 |

**Table 9 Time Series OOS Prediction Mean Accuracy across Models by Return State Transition 196301:201912**

Table 9 presents our OOS modeling prediction *average* accuracies of return state transitions from the old states to the new states across models. Specifically, we calculate the OOS prediction accuracies of each classification model and form a percentage accuracy table similar to the table below. We then average the numbers across all the models. A stock in return state 1 means that the stock delivers a return that is among the worst performing returns of the trading month. A stock in return state 10 indicates that the stock are among the stocks delivering the best performing returns of the trading month. Details of the return state definition can be found in Table 1.

Combining what is demonstrated in Table 8, Table 9 shows that our models benefit significantly from the most certain return states, i.e., return states 1 and 10. Our models almost give up the most uncertain states, i.e. return states 3, 4, and 9.

| | New 1 | New 2 | New 3 | New 4 | New 5 | New 6 | New 7 | New 8 | New 9 | New 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Old 1 | 0.5054 | 0.0228 | 0.0045 | 0.0020 | 0.0142 | 0.0446 | 0.0504 | 0.0476 | 0.0928 | 0.4449 |
| Old 2 | 0.4980 | 0.0735 | 0.0122 | 0.0064 | 0.0456 | 0.1478 | 0.1309 | 0.1040 | 0.1516 | 0.2424 |
| Old 3 | 0.4403 | 0.0902 | 0.0158 | 0.0125 | 0.0787 | 0.2515 | 0.1773 | 0.1052 | 0.1266 | 0.1753 |
| Old 4 | 0.4098 | 0.0874 | 0.0195 | 0.0146 | 0.0995 | 0.3184 | 0.2009 | 0.1023 | 0.0993 | 0.1475 |
| Old 5 | 0.4226 | 0.0832 | 0.0177 | 0.0152 | 0.1034 | 0.3350 | 0.2087 | 0.0923 | 0.0874 | 0.1404 |
| Old 6 | 0.4244 | 0.0822 | 0.0200 | 0.0153 | 0.1101 | 0.3358 | 0.2028 | 0.0963 | 0.0848 | 0.1390 |
| Old 7 | 0.4094 | 0.0942 | 0.0245 | 0.0159 | 0.1028 | 0.3333 | 0.2008 | 0.1025 | 0.0849 | 0.1303 |
| Old 8 | 0.4376 | 0.1093 | 0.0295 | 0.0189 | 0.1016 | 0.3133 | 0.1748 | 0.0991 | 0.0876 | 0.1293 |
| Old 9 | 0.5011 | 0.1472 | 0.0350 | 0.0221 | 0.0834 | 0.2463 | 0.1214 | 0.1033 | 0.0948 | 0.1270 |
| Old 10 | 0.7522 | 0.1173 | 0.0257 | 0.0166 | 0.0446 | 0.1067 | 0.0586 | 0.0613 | 0.0667 | 0.0992 |

**Table 10 OOS Prediction By-Class Accuracy 196301:201912**

This table summarizes the key statistical metrics that measures the performance of our classification models in the OOS predictions covering 196301:201912 by class for the two modeling architectures separately. Our models splits the big question of what is the return state of a stock in the next time period as smaller binary choice questions as describe in Section 2. Specifically, our models ask whether an observation will be in return state 1 or not, whether the observation will be in return state 2 or not, whether the observation will be in return state 3 or not and so on. In short, our models break down the multiclass classification problem into smaller binary classification problems. The evaluation of the accuracy with the statistical metrics by each return state can thus help us understand the modeling performance for each return state. For more details, section 2.2.2 discusses the details of the statistical metrics. Table 1 defines the return states. Table 2 describes model specification.

In the table below, prevalence is the percentage of the associated return state in the population. Sensitivity measures the proportion of correctly predicted positives, while specificity measures the proportion correctly predicted negatives. Sensitivity is also called recall. As an example, for the first line of return state 1, a positive means a prediction of 1 and a negative means a prediction of 0, where the prediction of 1 indicates that the return of the stock in the next period will be in return state 1 and the prediction of 0 indicates that the return of the stock in the next period will not be in return state 1. Precision measures the proportion of correct positives over the sum of correct positives and incorrect negatives. F1 is the balanced F score which is calculated as the harmonic mean of precision and recall. Balanced accuracy is the average of the accuracy between the prediction of positives and the prediction of negatives. In general, greater values in the metrics of sensitivity, specificity, precision, F1 and balanced accuracy indicate better performance.

Panel A presents the by-class metrics in term of average across the neural network models and Panel B presents the by-class metrics in term of average across the tree models. The two panels confirm our findings that our models perform well in return states with high certainty and almost give up the return states with high uncertainty. The balanced accuracy column also shows that our predictions balance the prediction between the positives and negatives.

Panel A: OOS Prediction Average of By-Class Metrics across ANN Models 196301:201912

| Return State | Prevalence | Sensitivity | Specificity | Precision | Recall | F1 | Balanced Accuracy |
|---:|---:|---:|---:|---:|---:|---:|---:|
| 1 | 0.0996 | 0.4865 | 0.7993 | 0.2146 | 0.4865 | 0.2948 | 0.6429 |
| 2 | 0.0996 | 0.0671 | 0.9482 | 0.1259 | 0.0671 | 0.0807 | 0.5076 |
| 3 | 0.0988 | 0.0298 | 0.9742 | 0.1139 | 0.0298 | 0.0391 | 0.5020 |
| 4 | 0.0984 | 0.0194 | 0.9836 | 0.1134 | 0.0194 | 0.0227 | 0.5015 |
| 5 | 0.0991 | 0.0898 | 0.9347 | 0.1122 | 0.0898 | NA | 0.5122 |
| 6 | 0.1008 | 0.2326 | 0.8413 | 0.1393 | 0.2326 | 0.1547 | 0.5370 |
| 7 | 0.1013 | 0.1994 | 0.8539 | 0.1329 | 0.1994 | 0.1417 | 0.5266 |
| 8 | 0.1016 | 0.0678 | 0.9436 | 0.1142 | 0.0678 | 0.0633 | 0.5057 |
| 9 | 0.1004 | 0.1238 | 0.8982 | NA | 0.1238 | NA | 0.5110 |
| 10 | 0.1003 | 0.2022 | 0.8807 | 0.1612 | 0.2022 | 0.1718 | 0.5414 |

Panel B: OOS Prediction Average of By-Class Metrics across Tree Models 196301:201912

| Return State | Prevalence | Sensitivity | Specificity | Precision | Recall | F1 | Balanced Accuracy |
|---:|---:|---:|---:|---:|---:|---:|---:|
| 1 | 0.0996 | 0.5332 | 0.7798 | 0.2118 | 0.5332 | 0.3027 | 0.6565 |
| 2 | 0.0996 | 0.1122 | 0.9132 | 0.1249 | 0.1122 | 0.1173 | 0.5127 |
| 3 | 0.0988 | 0.0136 | 0.9891 | 0.1226 | 0.0136 | 0.0227 | 0.5013 |
| 4 | 0.0984 | 0.0104 | 0.9914 | 0.1143 | 0.0104 | 0.0184 | 0.5009 |
| 5 | 0.0991 | 0.0782 | 0.9444 | 0.1330 | 0.0782 | 0.0968 | 0.5113 |
| 6 | 0.1008 | 0.2859 | 0.8153 | 0.1478 | 0.2859 | 0.1945 | 0.5506 |
| 7 | 0.1013 | 0.1301 | 0.9066 | 0.1358 | 0.1301 | 0.1316 | 0.5184 |
| 8 | 0.1016 | 0.1143 | 0.9069 | 0.1221 | 0.1143 | 0.1173 | 0.5106 |
| 9 | 0.1004 | 0.0775 | 0.9412 | 0.1281 | 0.0775 | 0.0939 | 0.5093 |
| 10 | 0.1003 | 0.2014 | 0.8740 | 0.1517 | 0.2014 | 0.1726 | 0.5377 |

(Note: the NAs are related to ANN3 128 which makes no prediction on return state 9 in the second half of our sample. We are updating the tables to report better metrics.)

**Table 11 Accuracy and Characteristics at the Individual Stock Level: GBM8 100 Example 196301:201912**

The table below presents a regression of OOS prediction accuracy on characteristics across individual stocks. The OOS prediction accuracy is produced by GBM8 100 over the entire sample coverage 196301:201912. For each stocks, we calculate the prediction accuracy during its existence in our sample and we include all 95 numeric characteristics augmented with the number of appearance (n) in the sample. The regression presents an R-squared of 0.485.

| Variable | Estimate | P Values | Variable | Estimate | P Values |
|---|---|---|---|---|---|
| ***Intangibles*** | | | | | |
| Dispersion in Forecasted EPS | -0.010 | **0.000** | Accrual Volatility | 0.003 | 0.131 |
| Number of Analysts Covering Stock | -0.005 | **0.000** | Current Ratio | -0.003 | 0.150 |
| Earnings Volatility | -0.004 | **0.000** | % Change Sales-to-Inventory | 0.002 | 0.160 |
| Cash Flow Volatility | -0.006 | **0.000** | Absolute Accruals | 0.001 | 0.395 |
| Secured Debt | -0.003 | **0.001** | Sales to Cash | -0.001 | 0.479 |
| Cash Holdings | -0.003 | **0.002** | % Change in Quick Ratio | 0.002 | 0.499 |
| Working Capital Accruals | 0.003 | **0.002** | Debt Capacity/Firm Tangibility | -0.001 | 0.509 |
| % Change in Sales - % Change in Inventory | -0.004 | **0.004** | Industry-Adjusted Change in Employees | 0.001 | 0.571 |
| % Change in Sales - % Change in SG&A | -0.003 | **0.020** | % Change in Sales - % Change in A/R | 0.000 | 0.633 |
| Employee Growth Rate | -0.003 | **0.025** | Sales to Receivables | 0.000 | 0.637 |
| # Years Since First Compustat Coverage | 0.002 | **0.042** | % Change in Gross Margin - % Change in Sales | 0.000 | 0.674 |
| Industry Sales Concentration | 0.001 | **0.046** | Quick Ratio | 0.001 | 0.680 |
| Sales to Inventory | 0.001 | **0.074** | Real Estate Holdings | 0.000 | 0.712 |
| R&D to Market Capitalization | -0.002 | **0.083** | Organizational Capital | 0.000 | 0.718 |
| Percent Accruals | -0.002 | **0.088** | % Change in Current Ratio | 0.000 | 0.915 |
| Growth in Long-Term Debt | 0.002 | **0.099** | R&D to Sales | 0.000 | 0.966 |
| Change in Number of Analysts | -0.005 | 0.113 | | | |
| ***Investment*** | | | | | |
| Growth in Long Term Net Operating Assets | -0.005 | **0.000** | Industry Adjusted % Change in Capital Expenditures | 0.001 | 0.222 |
| Asset Growth | 0.002 | **0.082** | Growth in Capital Expenditures | 0.000 | 0.713 |
| Corporate Investment | -0.003 | **0.099** | | | |
| ***Momentum*** | | | | | |
| 1-Month Momentum | -0.130 | **0.000** | Revenue Surprise | -0.004 | **0.002** |
| Industry Momentum | -0.007 | **0.000** | Change in Forecasted EPS | 0.009 | **0.028** |
| 12-Month Momentum | 0.027 | **0.000** | Change in 6-Month Momentum | -0.007 | **0.044** |
| Change in Tax Expense | 0.010 | **0.000** | Unexpected Quarterly Earnings | 0.004 | **0.049** |
| 36-Month Momentum | 0.011 | **0.000** | Number of Earnings Increases | 0.002 | **0.058** |
| Industry Return | 0.013 | **0.000** | Earnings Announcement Return | 0.004 | 0.117 |
| 6-Month Momentum | -0.015 | **0.000** | | | |
| ***Profitability*** | | | | | |
| Return on Equity | -0.009 | **0.000** | Industry-Adjusted Change in Profit Margin | -0.001 | 0.311 |
| Leverage | 0.004 | **0.000** | Industry-Adjusted Change in Asset Turnover | 0.001 | 0.351 |
| Return on Assets | -0.005 | **0.000** | Cash Productivity | -0.001 | 0.353 |
| Financial Statements Score | 0.003 | **0.011** | Gross Profitability | 0.001 | 0.458 |
| Return on Invested Capital | -0.002 | **0.014** | Tax Income to Book Income | 0.000 | 0.918 |
| Financial Statements Score | -0.001 | 0.218 | Operating Profitability | 0.000 | 0.940 |

**Table 11 (Continues)**

| Variable | Estimate | P Values | Variable | Estimate | P Values |
|---|---|---|---|---|---|
| ***Trading Frictions*** | | | | | |
| Beta | -0.042 | **0.000** | Price Delay | -0.006 | **0.000** |
| Illiquidity | -0.011 | **0.000** | Industry Adjusted Size | 0.004 | **0.000** |
| Abnormal Earnings Announcement Volume | -0.010 | **0.000** | Maximum Daily Return | -0.014 | **0.000** |
| Zero Trading Days | 0.006 | **0.000** | Volatility of Liquidity (Dollar Trading Volume) | 0.004 | **0.000** |
| Bid-Ask Spread | 0.012 | **0.000** | Share Turnover | -0.006 | **0.005** |
| Dollar Trading Volume | 0.017 | **0.000** | Volatility of Liquidity (Share Turnover) | 0.004 | 0.128 |
| Beta Squared | 0.037 | **0.000** | Idiosyncratic Return Volatility | 0.001 | 0.393 |
| Return Volatility | 0.069 | **0.000** | | | |
| ***Value vs. Growth*** | | | | | |
| Sales to Price | -0.009 | **0.000** | Capital Expenditures and Inventory | 0.002 | **0.060** |
| Dividend to Price | 0.006 | **0.000** | % Change in Depreciation | 0.002 | 0.143 |
| Forecasted Growth in 5-year EPS | -0.005 | **0.001** | Industry-Adjusted BM | 0.001 | 0.270 |
| Cash Flow to Debt | -0.003 | **0.001** | Cash Flow to Price Ratio | -0.001 | 0.484 |
| Book-to-Market | -0.003 | **0.001** | Change in Shares Outstanding | -0.001 | 0.513 |
| Growth in Common Shareholder Equity | -0.002 | **0.010** | Depreciation/PP&E | 0.000 | 0.622 |
| Change in Inventory | -0.003 | **0.029** | Earnings Announcement Return | -0.001 | 0.673 |
| Industry-Adjusted Cash Flow to Price RATIO | 0.003 | **0.029** | Scaled Earnings Forecast | 0.000 | 0.776 |
| Sales Growth | 0.003 | **0.058** | | | |
| ***Others*** | | | | | |
| Intercept | 0.162 | 0.000 | Appearance in Sample | 0.000 | 0.000 |

**Table 12 Model Certainty and Characteristics at the Individual Stock Level: GBM8 100 Example 196301:201912**

Table 12 presents the regression results of model certainty as measured by the variance of predicted probabilities. Higher probability variance means the model is able to distinguish high probability future states from the low probability future states. Therefore, the variance of predicted probabilities is a natural measure for model certainty. Higher variance means the model is more certain in general about the possible outcomes. Note that the model certainty is a pre-realization measure. The results below are from GBM8 100. We first calculate the variance of predicted probabilities of all 10 states for each stocks in each month. We then obtain the averages of the variance and the characteristics across the time for each stocks. Finally, we regress the variance on the characteristics. Note that the characteristics are normalized for each month across stocks such that the averages of the characteristics can reflect the relative position of a stock along the characteristics across the time. The regression presents an R-squared of 0.673.

| Variable | Estimate | P Values | Variable | Estimate | P Values |
|---|---|---|---|---|---|
| | | | ***Intangibles*** | | |
| Dispersion in Forecasted EPS | -0.0003 | **0.000** | Debt Capacity/Firm Tangibility | 0.000 | **0.034** |
| Cash Holdings | -0.0001 | **0.000** | Industry Sales Concentration | 0.000 | **0.050** |
| Secured Debt | -0.0001 | **0.000** | Number of Analysts Covering Stock | 0.000 | **0.054** |
| % Change Sales-to-Inventory | 0.0001 | **0.000** | Sales to Cash | 0.000 | **0.067** |
| # Years Since First Compustat Coverage | 0.0001 | **0.000** | Quick Ratio | 0.000 | 0.165 |
| % Change in Sales - % Change in Inventory | -0.0001 | **0.000** | % Change in Sales - % Change in A/R | 0.000 | 0.202 |
| Sales to Inventory | 0.0000 | **0.000** | R&D to Market Capitalization | 0.000 | 0.224 |
| Earnings Volatility | 0.0000 | **0.001** | Real Estate Holdings | 0.000 | 0.289 |
| Change in Number of Analysts | -0.0001 | **0.002** | % Change in Sales - % Change in SG&A | 0.000 | 0.590 |
| Employee Growth Rate | -0.0001 | **0.003** | Cash Flow Volatility | 0.000 | 0.604 |
| Current Ratio | -0.0001 | **0.003** | Percent Accruals | 0.000 | 0.678 |
| R&D to Sales | 0.0000 | **0.004** | Accrual Volatility | 0.000 | 0.811 |
| Organizational Capital | 0.0000 | **0.007** | % Change in Gross Margin - % Change in Sales | 0.000 | 0.847 |
| Working Capital Accruals | 0.0000 | **0.008** | Growth in Long-Term Debt | 0.000 | 0.945 |
| Sales to Receivables | 0.0000 | **0.011** | % Change in Current Ratio | 0.000 | 0.949 |
| Absolute Accruals | 0.0000 | **0.013** | % Change in Quick Ratio | 0.000 | 0.995 |
| Industry-Adjusted Change in Employees | 0.0000 | **0.026** | | | |
| | | | ***Investment*** | | |
| Asset Growth | 0.0000 | **0.011** | Industry Adjusted % Change in Capital Expenditures | 0.000 | 0.703 |
| Growth in Long Term Net Operating Assets | 0.0000 | 0.212 | Corporate Investment | 0.000 | 0.950 |
| Growth in Capital Expenditures | 0.0000 | 0.444 | | | |
| | | | ***Momentum*** | | |
| 6-Month Momentum | -0.0006 | **0.000** | 12-Month Momentum | -0.0001 | **0.005** |
| 1-Month Momentum | -0.0005 | **0.000** | Industry Momentum | 0.0000 | **0.055** |
| Change in Tax Expense | 0.0001 | **0.000** | Number of Earnings Increases | 0.0000 | 0.223 |
| Change in Forecasted EPS | 0.0003 | **0.000** | Unexpected Quarterly Earnings | 0.0000 | 0.430 |
| 36-Month Momentum | 0.0003 | **0.000** | Earnings Announcement Return | 0.0000 | 0.906 |
| Industry Return | 0.0004 | **0.000** | Change in 6-Month Momentum | 0.0000 | 0.988 |
| Revenue Surprise | -0.0001 | **0.000** | | | |

**Table 12 (Continues)**

| Variable | Estimate | P Values | Variable | Estimate | P Values |
|---|---|---|---|---|---|
| ***Profitability*** | | | | | |
| Return on Assets | -0.0002 | 0.000 | Tax Income to Book Income | 0.0000 | 0.007 |
| Return on Equity | -0.0001 | **0.000** | Financial Statements Score | 0.0000 | **0.059** |
| Leverage | 0.0002 | **0.000** | Industry-Adjusted Change in Asset Turnover | 0.0000 | 0.120 |
| Cash Productivity | 0.0000 | **0.001** | Industry-Adjusted Change in Profit Margin | 0.0000 | 0.303 |
| Return on Invested Capital | 0.0000 | **0.001** | Operating Profitability | 0.0000 | 0.400 |
| Gross Profitability | 0.0000 | **0.006** | Financial Statements Score | 0.0000 | 0.691 |
| ***Trading Frictions*** | | | | | |
| Beta | -0.0011 | **0.000** | Volatility of Liquidity (Dollar Trading Volume) | 0.0003 | **0.000** |
| Abnormal Earnings Announcement Volume | -0.0003 | **0.000** | Bid-Ask Spread | 0.0006 | **0.000** |
| Price Delay | -0.0002 | **0.000** | Maximum Daily Return | 0.0006 | **0.000** |
| Illiquidity | -0.0002 | **0.000** | Beta Squared | 0.0009 | **0.000** |
| Zero Trading Days | 0.0001 | **0.000** | Return Volatility | 0.0002 | **0.000** |
| Industry Adjusted Size | 0.0002 | **0.000** | Share Turnover | -0.0001 | **0.022** |
| Idiosyncratic Return Volatility | 0.0002 | **0.000** | Volatility of Liquidity (Share Turnover) | 0.0001 | **0.022** |
| Dollar Trading Volume | 0.0003 | **0.000** | | | |
| ***Value vs. Growth*** | | | | | |
| Sales to Price | -0.0003 | **0.000** | Capital Expenditures and Inventory | -0.0001 | **0.001** |
| Forecasted Growth in 5-year EPS | -0.0002 | **0.000** | Growth in Common Shareholder Equity | 0.0000 | **0.004** |
| Cash Flow to Price Ratio | -0.0001 | **0.000** | % Change in Depreciation | 0.0000 | **0.006** |
| Industry-Adjusted BM | -0.0001 | **0.000** | Change in Inventory | 0.0000 | **0.079** |
| Book-to-Market | -0.0001 | **0.000** | Earnings Announcement Return | 0.0000 | 0.158 |
| Cash Flow to Debt | -0.0001 | **0.000** | Change in Shares Outstanding | 0.0000 | 0.562 |
| Depreciation/PP&E | -0.0001 | **0.000** | Industry-Adjusted Cash Flow to Price RATIO | 0.0000 | 0.690 |
| Dividend to Price | 0.0003 | **0.000** | Sales Growth | 0.0000 | 0.905 |
| Scaled Earnings Forecast | -0.0001 | **0.000** | | | |
| ***Others*** | | | | | |
| Intercept | 0.0014 | **0.000** | Appearance in Sample | 0.0000 | **0.000** |

**Table 13 Accuracy, Model Certainty and Characteristics in Categories: GBM8 100 Example**

Table 13 summarizes the analysis of relation related to modeling characteristics in categories following the categorization of Hou et al. 2019. Panel A summarizes the information about the relation between OOS prediction accuracy and the characteristics. Panel B summarizes the information about the relation between model certainty and the characteristics. The Table 13 is based on Table 11 and Table 12. The categorization details are included in Tables 11 and 12.

Panel A: OOS Prediction Accuracy on Characteristics

| Variable Category | Sum of Significant Effect | Number of Significant Effects | Total Number of Variables |
|---|---:|---:|---:|
| Intangibles | -0.0367 | 16 | 33 |
| Investment | -0.0054 | 3 | 5 |
| Momentum | -0.0847 | 12 | 13 |
| Profitability | -0.0091 | 5 | 12 |
| Trading Frictions | 0.0578 | 13 | 15 |
| Value vs. Growth | -0.0102 | 10 | 17 |
| Others | 0.0000 | 2 | 2 |

Panel B: Model Certainty on Characteristics

| Variable Category | Sum of Significant Effect | Number of Significant Effects | Total Number of Variables |
|---|---:|---:|---:|
| Intangibles | -0.0007 | 21 | 33 |
| Investment | 0.0000 | 1 | 5 |
| Momentum | -0.0001 | 9 | 13 |
| Profitability | -0.0002 | 8 | 12 |
| Trading Frictions | 0.0017 | 15 | 15 |
| Value vs. Growth | -0.0007 | 13 | 17 |
| Others | 0.0014 | 2 | 2 |

**Table 14 OOS Rolling Window Relation: GBM8 100 Example 199201:201912**

The table below presents the rolling correlation between each pair of value weighted market return, model confidence, model certainty and OOS prediction accuracy. We measure model confidence as the maximum predicted return state probability. The definition of model certainty is the same as in Table 12. For model confidence and model certainty, we first calculate the values at the individual stock level and then we take the averages across the stocks in the month. The correlation calculation is based on normalized scales across the time period from 1992 to 2019. Model confidence, model certainty and model accuracy are all from GBM8 100.

| Rolling Window | MKT<br>and<br>Model Confidence | MKT<br>and<br>Model Certainty | Accuracy<br>And<br>Model Confidence | Accuracy<br>and<br>Model Certainty | MKT<br>and<br>Accuracy |
|---|---|---|---|---|---|
| 1 | -0.063 | 0.019 | 0.224 | 0.199 | 0.152 |
| 6 | -0.273 | -0.010 | 0.401 | 0.321 | 0.064 |
| 12 | -0.400 | -0.101 | 0.440 | 0.349 | 0.036 |
| 18 | -0.440 | -0.124 | 0.450 | 0.382 | 0.070 |
| 24 | -0.431 | -0.115 | 0.457 | 0.407 | 0.120 |
| 30 | -0.420 | -0.107 | 0.467 | 0.429 | 0.172 |
| 36 | -0.399 | -0.107 | 0.470 | 0.442 | 0.235 |
| 42 | -0.382 | -0.113 | 0.477 | 0.459 | 0.298 |
| 48 | -0.366 | -0.119 | 0.491 | 0.484 | 0.339 |
| 54 | -0.356 | -0.117 | 0.514 | 0.515 | 0.367 |
| 60 | -0.349 | -0.112 | 0.542 | 0.549 | 0.390 |
| 66 | -0.341 | -0.106 | 0.575 | 0.585 | 0.423 |
| 72 | -0.332 | -0.092 | 0.604 | 0.614 | 0.454 |
| 78 | -0.322 | -0.072 | 0.627 | 0.638 | 0.461 |
| 84 | -0.322 | -0.069 | 0.646 | 0.660 | 0.453 |
| 90 | -0.329 | -0.068 | 0.664 | 0.678 | 0.424 |
| 96 | -0.334 | -0.065 | 0.682 | 0.691 | 0.381 |
| 102 | -0.321 | -0.045 | 0.700 | 0.699 | 0.345 |
| 108 | -0.293 | 0.011 | 0.720 | 0.708 | 0.324 |
| 114 | -0.258 | 0.052 | 0.737 | 0.713 | 0.324 |
| 120 | -0.218 | 0.089 | 0.751 | 0.714 | 0.337 |
| 240 | 0.284 | -0.697 | -0.069 | -0.938 | 0.812 |

**Table 15 Probability Adjusted Portfolio Allocation: Boosting the Performance 1963:2019**

Predictions of the machine learning models are subjective. Therefore, models may have systematic bias in favoring certain outcomes when we classify the future return state based on the maximum predicted probability. An alternative to avoid this systematic bias in favoring certain outcomes is to use relative ranking of the predicted probabilities for each possible return state. The table below demonstrates a drastic increase in performance when we form portfolios utilizing the relative positions of predicted return state probabilities for each month. Specifically, we hold long positions in stocks with predicted return state 10 probability in the top decile of the month and we require the stocks to not be in the top deciles of the predicted return state probabilities for return state 1-9. Similarly, we short stocks with predicted return state 1 probability in the top decile of the month and we require the stocks to not be in the top deciles of the predicted return state probabilities for return state 2-10. At the same time, we adjust the weights of the stocks in the long and the short legs with their predicted return state 10 probabilities and their predicted return state 1 probabilities. In other words, among the selected stocks based on the relative predicted return state probabilities, we allocate more positions to stocks that have higher predicted probabilities of ending up in return state 1 and return state 10. Comparing to Table 6, the long-short portfolios experience drastic performance increase after the probability based weight adjustment. The model, ANN3 128, which fails in Table 4 and our figures for economic performance are now delivering a high performance.

Panel A: Probability Adjusted Equal Weight Long-Short Portfolio Economic Performance

| Model | Mean | SD | Skewness | SR | CEQ | Cumulative Return | Min | Max DD |
|---|---:|---:|---:|---:|---:|---:|---:|---:|
| Buy-Hold | 0.01 | 0.06 | -0.19 | 0.13 | 0.01 | 5478% | -0.28 | 0.73 |
| ANN1 128 | 0.03 | 0.05 | 0.24 | 0.54 | 0.02 | 1747545888% | -0.24 | 0.40 |
| ANN1 16 | 0.03 | 0.05 | 1.46 | 0.74 | 0.03 | 537367902806% | -0.17 | 0.23 |
| ANN1 32 | 0.03 | 0.04 | 0.43 | 0.80 | 0.03 | 585717134036% | -0.18 | 0.28 |
| ANN1 64 | 0.03 | 0.04 | 0.67 | 0.79 | 0.03 | 126545896631% | -0.09 | 0.26 |
| ANN2 128 | 0.03 | 0.07 | -2.09 | 0.37 | 0.02 | 6384408% | -0.71 | 0.77 |
| ANN2 32 | 0.05 | 0.09 | 0.42 | 0.61 | 0.05 | 53503237666690300% | -0.51 | 0.75 |
| ANN2 64 | 0.04 | 0.05 | 0.06 | 0.84 | 0.04 | 21946520% | -0.13 | 0.16 |
| **ANN3 128** | **0.04** | **0.06** | **-0.27** | **0.64** | **0.04** | **27745008%** | **-0.28** | **0.38** |
| ANN3 64 | 0.03 | 0.05 | -0.20 | 0.60 | 0.03 | 116318933473% | -0.41 | 0.56 |
| ANN4 128 | 0.04 | 0.05 | 0.13 | 0.73 | 0.03 | 857535837585% | -0.23 | 0.41 |
| DART2 100 | 0.02 | 0.04 | 0.11 | 0.45 | 0.02 | 9780805% | -0.26 | 0.35 |
| DART4 100 | 0.03 | 0.05 | 1.17 | 0.59 | 0.03 | 6142361836% | -0.20 | 0.29 |
| DART6 100 | 0.04 | 0.05 | 1.74 | 0.81 | 0.04 | 4337282808905% | -0.13 | 0.14 |
| DART8 100 | 0.04 | 0.05 | 1.61 | 0.89 | 0.04 | 22990589725743% | -0.08 | 0.13 |
| DRF2 200 | 0.02 | 0.04 | -0.41 | 0.49 | 0.02 | 117805370% | -0.24 | 0.42 |
| DRF4 200 | 0.03 | 0.05 | 0.36 | 0.69 | 0.03 | 254014955988% | -0.22 | 0.26 |
| DRF6 200 | 0.04 | 0.05 | 0.54 | 0.73 | 0.04 | 3527074114759% | -0.17 | 0.23 |
| DRF8 200 | 0.04 | 0.05 | 0.66 | 0.82 | 0.04 | 51326803125824% | -0.17 | 0.25 |
| GBM2 100 | 0.02 | 0.05 | 0.33 | 0.49 | 0.02 | 152137336% | -0.18 | 0.31 |
| GBM4 100 | 0.04 | 0.05 | 0.52 | 0.82 | 0.04 | 9280507544099% | -0.13 | 0.19 |
| GBM6 100 | 0.05 | 0.05 | 1.03 | 0.95 | 0.05 | 4697009163630370% | -0.08 | 0.10 |
| GBM8 100 | 0.05 | 0.05 | 1.16 | 0.96 | 0.05 | 9480038488767790% | -0.07 | 0.18 |

**Table 15 (Continues)**

| Panel B: Probability Adjusted Value Weight Long-Short Portfolio Economic Performance | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Mean | SD | Skewness | SR | CEQ | Cumulative Return | Min | Max DD |
| Buy-Hold | 0.01 | 0.04 | -0.56 | 0.12 | 0.00 | 2004% | -0.23 | 0.55 |
| ANN1 128 | 0.02 | 0.05 | -0.12 | 0.34 | 0.02 | 11945601% | -0.28 | 0.57 |
| ANN1 16 | 0.02 | 0.06 | 1.87 | 0.41 | 0.02 | 197771788% | -0.20 | 0.74 |
| ANN1 32 | 0.03 | 0.05 | 0.31 | 0.47 | 0.02 | 950511246% | -0.27 | 0.36 |
| ANN1 64 | 0.02 | 0.05 | 0.21 | 0.42 | 0.02 | 124348450% | -0.20 | 0.68 |
| ANN2 128 | 0.02 | 0.10 | 0.87 | 0.18 | 0.01 | 640963% | -0.46 | 0.85 |
| ANN2 32 | 0.03 | 0.06 | 0.50 | 0.52 | 0.03 | 66744428793% | -0.26 | 0.41 |
| ANN2 64 | 0.03 | 0.12 | 0.92 | 0.22 | 0.02 | 65858765% | -0.57 | 0.86 |
| **ANN3 128** | **0.03** | **0.04** | **0.29** | **0.64** | **0.03** | **1239601%** | **-0.09** | **0.14** |
| ANN3 64 | 0.02 | 0.05 | -0.44 | 0.44 | 0.02 | 269369208% | -0.33 | 0.39 |
| ANN4 128 | 0.03 | 0.06 | -0.64 | 0.44 | 0.02 | 1554362544% | -0.40 | 0.44 |
| DART2 100 | 0.02 | 0.06 | -0.64 | 0.32 | 0.02 | 6500948% | -0.36 | 0.54 |
| DART4 100 | 0.02 | 0.06 | 0.46 | 0.30 | 0.02 | 3974018% | -0.28 | 0.47 |
| DART6 100 | 0.03 | 0.05 | 1.04 | 0.51 | 0.03 | 2934823199% | -0.23 | 0.44 |
| DART8 100 | 0.03 | 0.06 | -0.04 | 0.45 | 0.03 | 2390819465% | -0.26 | 0.36 |
| DRF2 200 | 0.01 | 0.06 | -0.39 | 0.23 | 0.01 | 192536% | -0.27 | 0.51 |
| DRF4 200 | 0.02 | 0.06 | -0.04 | 0.33 | 0.02 | 23122352% | -0.26 | 0.65 |
| DRF6 200 | 0.02 | 0.06 | 0.61 | 0.35 | 0.02 | 99471739% | -0.20 | 0.65 |
| DRF8 200 | 0.03 | 0.06 | 0.25 | 0.46 | 0.03 | 3018302657% | -0.20 | 0.56 |
| GBM2 100 | 0.02 | 0.06 | -0.55 | 0.28 | 0.02 | 3567034% | -0.31 | 0.67 |
| GBM4 100 | 0.03 | 0.06 | 0.12 | 0.43 | 0.03 | 3229834491% | -0.30 | 0.45 |
| GBM6 100 | 0.03 | 0.06 | 0.16 | 0.50 | 0.03 | 31495600068% | -0.25 | 0.37 |
| GBM8 100 | 0.03 | 0.06 | 0.04 | 0.53 | 0.03 | 51378487503% | -0.30 | 0.35 |

**Table 16 Time Series IS Training Accuracy and Binomial Test**

Table 16 presents the overall accuracy of the predictions on IS training set across model specifications for the time period from 196301:201912. The setup is the same as Table 7. Panel A shows overall accuracy of training set across models for 199201:201912. Panel B shows the same information for 199201:201912.

Panel A: Overall Accuracy of Training Set Across Models 196301:199112

| Model | Accuracy | Kappa | Lower 99% Bound | Upper 99% Bound | No Info Accuracy | No Info P Value | Martingale Accuracy | Martingale P Value |
|---|---|---|---|---|---|---|---|---|
| ANN1 16 | 0.156 | 0.062 | 0.155 | 0.157 | 0.103 | **0.000** | 0.113 | **0.000** |
| ANN1 32 | 0.154 | 0.059 | 0.154 | 0.155 | 0.103 | **0.000** | 0.113 | **0.000** |
| ANN1 64 | 0.152 | 0.057 | 0.151 | 0.152 | 0.103 | **0.000** | 0.113 | **0.000** |
| ANN1 128 | 0.151 | 0.056 | 0.150 | 0.152 | 0.103 | **0.000** | 0.113 | **0.000** |
| ANN2 32 | 0.153 | 0.058 | 0.152 | 0.154 | 0.103 | **0.000** | 0.113 | **0.000** |
| ANN2 64 | 0.152 | 0.058 | 0.152 | 0.153 | 0.103 | **0.000** | 0.113 | **0.000** |
| ANN2 128 | 0.151 | 0.056 | 0.150 | 0.152 | 0.103 | **0.000** | 0.113 | **0.000** |
| ANN3 64 | 0.154 | 0.059 | 0.153 | 0.155 | 0.103 | **0.000** | 0.113 | **0.000** |
| ANN3 128 | 0.149 | 0.053 | 0.148 | 0.150 | 0.103 | **0.000** | 0.113 | **0.000** |
| ANN4 128 | 0.150 | 0.056 | 0.149 | 0.151 | 0.103 | **0.000** | 0.113 | **0.000** |
| DART2 100 | 0.153 | 0.058 | 0.152 | 0.153 | 0.103 | **0.000** | 0.113 | **0.000** |
| DART4 100 | 0.162 | 0.068 | 0.161 | 0.163 | 0.103 | **0.000** | 0.113 | **0.000** |
| DART6 100 | 0.173 | 0.081 | 0.172 | 0.174 | 0.103 | **0.000** | 0.113 | **0.000** |
| DART8 100 | 0.200 | 0.110 | 0.199 | 0.201 | 0.103 | **0.000** | 0.113 | **0.000** |
| DRF2 200 | 0.151 | 0.056 | 0.151 | 0.152 | 0.103 | **0.000** | 0.113 | **0.000** |
| DRF4 200 | 0.157 | 0.063 | 0.156 | 0.158 | 0.103 | **0.000** | 0.113 | **0.000** |
| DRF6 200 | 0.164 | 0.070 | 0.163 | 0.165 | 0.103 | **0.000** | 0.113 | **0.000** |
| DRF8 200 | 0.179 | 0.087 | 0.178 | 0.180 | 0.103 | **0.000** | 0.113 | **0.000** |
| GBM2 100 | 0.157 | 0.062 | 0.156 | 0.158 | 0.103 | **0.000** | 0.113 | **0.000** |
| GBM4 100 | 0.165 | 0.072 | 0.165 | 0.166 | 0.103 | **0.000** | 0.113 | **0.000** |
| GBM6 100 | 0.180 | 0.088 | 0.179 | 0.181 | 0.103 | **0.000** | 0.113 | **0.000** |
| GBM8 100 | 0.210 | 0.122 | 0.210 | 0.211 | 0.103 | **0.000** | 0.113 | **0.000** |

**Table 16 (Continues)**

| Panel B: Overall Accuracy of Training Set Across Models 199201:201912 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Accuracy | Kappa | Lower 99% Bound | Upper 99% Bound | No Info Accuracy | No Info P Value | Martingale Accuracy | Martingale P Value |
| ANN1 16 | 0.161 | 0.068 | 0.160 | 0.162 | 0.120 | **0.000** | 0.120 | **0.000** |
| ANN1 32 | 0.159 | 0.065 | 0.158 | 0.159 | 0.120 | **0.000** | 0.120 | **0.000** |
| ANN1 64 | 0.163 | 0.069 | 0.162 | 0.163 | 0.120 | **0.000** | 0.120 | **0.000** |
| ANN1 128 | 0.157 | 0.063 | 0.156 | 0.158 | 0.120 | **0.000** | 0.120 | **0.000** |
| ANN2 32 | 0.158 | 0.064 | 0.157 | 0.159 | 0.120 | **0.000** | 0.120 | **0.000** |
| ANN2 64 | 0.155 | 0.061 | 0.154 | 0.156 | 0.120 | **0.000** | 0.120 | **0.000** |
| ANN2 128 | 0.157 | 0.063 | 0.156 | 0.157 | 0.120 | **0.000** | 0.120 | **0.000** |
| ANN3 64 | 0.158 | 0.064 | 0.157 | 0.158 | 0.120 | **0.000** | 0.120 | **0.000** |
| ANN3 128 | 0.160 | 0.067 | 0.160 | 0.161 | 0.120 | **0.000** | 0.120 | **0.000** |
| ANN4 128 | 0.158 | 0.064 | 0.157 | 0.158 | 0.120 | **0.000** | 0.120 | **0.000** |
| DART2 100 | 0.159 | 0.066 | 0.159 | 0.160 | 0.120 | **0.000** | 0.120 | **0.000** |
| DART4 100 | 0.165 | 0.072 | 0.164 | 0.165 | 0.120 | **0.000** | 0.120 | **0.000** |
| DART6 100 | 0.174 | 0.082 | 0.173 | 0.175 | 0.120 | **0.000** | 0.120 | **0.000** |
| DART8 100 | 0.196 | 0.107 | 0.196 | 0.197 | 0.120 | **0.000** | 0.120 | **0.000** |
| DRF2 200 | 0.158 | 0.065 | 0.158 | 0.159 | 0.120 | **0.000** | 0.120 | **0.000** |
| DRF4 200 | 0.162 | 0.069 | 0.162 | 0.163 | 0.120 | **0.000** | 0.120 | **0.000** |
| DRF6 200 | 0.167 | 0.075 | 0.167 | 0.168 | 0.120 | **0.000** | 0.120 | **0.000** |
| DRF8 200 | 0.180 | 0.088 | 0.179 | 0.180 | 0.120 | **0.000** | 0.120 | **0.000** |
| GBM2 100 | 0.162 | 0.069 | 0.161 | 0.163 | 0.120 | **0.000** | 0.120 | **0.000** |
| GBM4 100 | 0.169 | 0.076 | 0.168 | 0.170 | 0.120 | **0.000** | 0.120 | **0.000** |
| GBM6 100 | 0.180 | 0.089 | 0.179 | 0.181 | 0.120 | **0.000** | 0.120 | **0.000** |
| GBM8 100 | 0.207 | 0.119 | 0.206 | 0.208 | 0.120 | **0.000** | 0.120 | **0.000** |

**Table 17 Cross-Sectional OOS Test Accuracy with Even Number Months 196302:201912**

This table presents the overall accuracy of the OOS predictions with even number months across model specifications for our cross-sectional evaluation. The sample split is demonstrated in Table 5 and the table setup is the same as Table 7. We confirm that our models are superior in cross-sectional OOS evaluation than the no information accuracy implied by the efficient market hypothesis. To illustrate the variable contribution with our cross-sectional setup, our results below are based the data with macro-level components.

| Model | Accuracy | Kappa | 95% Lower Bound | 95% Upper Bound | No Information Accuracy | P Value |
|---|---|---|---|---|---|---|
| ANN1 128 | 0.1476 | 0.0529 | 0.1470 | 0.1481 | 0.1019 | **0.0000** |
| ANN1 16 | 0.1503 | 0.0559 | 0.1497 | 0.1508 | 0.1019 | **0.0000** |
| ANN1 32 | 0.1499 | 0.0556 | 0.1494 | 0.1505 | 0.1019 | **0.0000** |
| ANN1 64 | 0.1543 | 0.0600 | 0.1538 | 0.1548 | 0.1019 | **0.0000** |
| ANN2 128 | 0.1507 | 0.0566 | 0.1501 | 0.1512 | 0.1019 | **0.0000** |
| ANN2 32 | 0.1491 | 0.0544 | 0.1485 | 0.1496 | 0.1019 | **0.0000** |
| ANN2 64 | 0.1539 | 0.0600 | 0.1534 | 0.1545 | 0.1019 | **0.0000** |
| ANN3 128 | 0.1544 | 0.0602 | 0.1539 | 0.1549 | 0.1019 | **0.0000** |
| ANN3 64 | 0.1535 | 0.0593 | 0.1530 | 0.1540 | 0.1019 | **0.0000** |
| ANN4 128 | 0.1516 | 0.0577 | 0.1511 | 0.1522 | 0.1019 | **0.0000** |
| DART2 100 | 0.1545 | 0.0605 | 0.1540 | 0.1551 | 0.1019 | **0.0000** |
| DART4 100 | 0.1579 | 0.0642 | 0.1574 | 0.1585 | 0.1019 | **0.0000** |
| DART6 100 | 0.1601 | 0.0666 | 0.1595 | 0.1606 | 0.1019 | **0.0000** |
| DART8 100 | 0.1608 | 0.0674 | 0.1602 | 0.1614 | 0.1019 | **0.0000** |
| DRF2 200 | 0.1558 | 0.0616 | 0.1552 | 0.1563 | 0.1019 | **0.0000** |
| DRF4 200 | 0.1577 | 0.0637 | 0.1571 | 0.1582 | 0.1019 | **0.0000** |
| DRF6 200 | 0.1594 | 0.0656 | 0.1588 | 0.1599 | 0.1019 | **0.0000** |
| DRF8 200 | 0.1598 | 0.0661 | 0.1592 | 0.1603 | 0.1019 | **0.0000** |
| GBM2 100 | 0.1582 | 0.0644 | 0.1577 | 0.1588 | 0.1019 | **0.0000** |
| GBM4 100 | 0.1600 | 0.0664 | 0.1595 | 0.1606 | 0.1019 | **0.0000** |
| GBM6 100 | 0.1609 | 0.0674 | 0.1603 | 0.1614 | 0.1019 | **0.0000** |
| GBM8 100 | 0.1614 | 0.0680 | 0.1608 | 0.1619 | 0.1019 | **0.0000** |

(Note: we are updating cross-sectional tables and new results should be available soon.)

**Table 18 CS Training Model Average Variable Importance**

Table 18 presents the variable importance of the cross-sectional models based on the training process covering 196301:201911. The table is based on our data including macro-level data components. The variable importance is calculated as the total sum of squared error that the associated variable is able to reduce during the model training process. The sample splits is defined in Table 5. To demonstrate the information in a concise way, we average across the models for the 2 main modeling architectures, ANN and trees, separately as demonstrated respectively in Panel A and B. Specifically, the trading related variables such as lagged idiosyncratic volatility play a part in our models. However, corporate announcement variables such as IPO history in the past 12 months, earning-price ratio, dividend-price ratio and other public information such as number of analysts and SIC classification all play important roles in our models. In addition, historical macro variables make marginal contribution to our models. Note that the suffix number indicates a specific category of the associated indicator variable. For example, ipo.0 stands for the no ipo indicator and sich.-1 stands for the indicator of no SIC information. Because of the difference in encoding the categorical variables (indicators) in our neural network models vs our tree models, our two architectures handle indicator variables differently.

Panel A: Top 50 Average Variable Importance Across Cross Sectional ANN Training Models

| Importance | Variable | Relative Importance | Scaled Importance | Percentage | Rank |
|---:|---|---:|---:|---:|---:|
| 1 | idiovol | 0.9968 | 0.9968 | 0.0184 | 1.1 |
| 2 | baspread | 0.7669 | 0.7669 | 0.0137 | 3.3 |
| 3 | sich2.60 | 0.7549 | 0.7549 | 0.0136 | 3.5 |
| 4 | retvol | 0.6986 | 0.6986 | 0.0125 | 3.8 |
| 5 | ipo.0 | 0.6538 | 0.6538 | 0.0115 | 4.5 |
| 6 | ipo.1 | 0.4564 | 0.4564 | 0.0078 | 8.9 |
| 7 | label10.0 | 0.4035 | 0.4035 | 0.0071 | 10.3 |
| 8 | label10.9 | 0.3975 | 0.3975 | 0.0069 | 11.6 |
| 9 | zerotrade | 0.3773 | 0.3773 | 0.0068 | 13.5 |
| 10 | sin.0 | 0.4056 | 0.4056 | 0.007 | 14.5 |
| 11 | mom12m | 0.3634 | 0.3634 | 0.0064 | 15.2 |
| 12 | sich2.49 | 0.372 | 0.372 | 0.0067 | 16.1 |
| 13 | mom1m | 0.3564 | 0.3564 | 0.0063 | 16.5 |
| 14 | dolvol | 0.363 | 0.363 | 0.0063 | 17.6 |
| 15 | sich2.10 | 0.3539 | 0.3539 | 0.0063 | 18.1 |
| 16 | mom6m | 0.3252 | 0.3252 | 0.0057 | 20.2 |
| 17 | beta | 0.3298 | 0.3298 | 0.0059 | 21.1 |
| 18 | sich2.-1 | 0.3138 | 0.3138 | 0.0055 | 21.3 |
| 19 | securedind.1 | 0.3244 | 0.3244 | 0.0055 | 23.6 |
| 20 | securedind.0 | 0.3103 | 0.3103 | 0.0053 | 27.9 |
| 21 | convind.0 | 0.3086 | 0.3086 | 0.0053 | 28.6 |
| 22 | dp | 0.3107 | 0.3107 | 0.0055 | 28.8 |
| 23 | rd.1 | 0.2967 | 0.2967 | 0.0051 | 29.6 |
| 24 | label10.8 | 0.3003 | 0.3003 | 0.005 | 31.6 |
| 25 | label10.1 | 0.2789 | 0.2789 | 0.0048 | 34.1 |
| 26 | label10.2 | 0.2847 | 0.2847 | 0.0048 | 34.1 |
| 27 | rd.-1 | 0.2788 | 0.2788 | 0.0048 | 35.3 |
| 28 | label10.7 | 0.273 | 0.273 | 0.0047 | 35.7 |
| 29 | divi.-1 | 0.2818 | 0.2818 | 0.0048 | 36.1 |
| 30 | divi.0 | 0.2839 | 0.2839 | 0.0049 | 36.4 |
| 31 | age | 0.2551 | 0.2551 | 0.0045 | 37.1 |
| 32 | sich2.56 | 0.2515 | 0.2515 | 0.0045 | 37.1 |
| 33 | sich2.28 | 0.2713 | 0.2713 | 0.0046 | 37.2 |
| 34 | turn | 0.261 | 0.261 | 0.0045 | 37.4 |
| 35 | divo.0 | 0.2844 | 0.2844 | 0.0049 | 37.6 |
| 36 | convind.1 | 0.2764 | 0.2764 | 0.0047 | 37.7 |
| 37 | label10.4 | 0.2692 | 0.2692 | 0.0046 | 38.4 |
| 38 | label10.6 | 0.2679 | 0.2679 | 0.0046 | 38.5 |
| 39 | label10.3 | 0.2682 | 0.2682 | 0.0046 | 38.9 |
| 40 | divo.-1 | 0.2683 | 0.2683 | 0.0046 | 40.2 |
| 41 | rd.0 | 0.2688 | 0.2688 | 0.0046 | 41 |
| 42 | invest.y | 0.2508 | 0.2508 | 0.0043 | 42.7 |
| 43 | sich2.63 | 0.2359 | 0.2359 | 0.0041 | 42.9 |
| 44 | sich2.13 | 0.2292 | 0.2292 | 0.004 | 51 |

Panel A: Top 50 Average Variable Importance Across Cross Sectional ANN Training Models **(Continues)**

| Importance | Variable | Relative Importance | Scaled Importance | Percentage | Rank |
|---:|:---|---:|---:|---:|---:|
| 45 | dy | 0.2186 | 0.2186 | 0.0039 | 51.2 |
| 46 | sich2.20 | 0.2172 | 0.2172 | 0.0038 | 51.6 |
| 47 | securedind.-1 | 0.2367 | 0.2367 | 0.0041 | 52.7 |
| 48 | ppicmm | 0.215 | 0.215 | 0.0037 | 53.5 |
| 49 | std_turn | 0.2073 | 0.2073 | 0.0037 | 55 |
| 50 | label10.5 | 0.2397 | 0.2397 | 0.0041 | 55.6 |

**Table 18 (Continues)**

| Panel B: Top 50 Average Variable Importance Across Cross Sectional Tree Training Models | | | | | |
|---|---|---|---|---|---|
| Importance | Variable | Relative Importance | Scaled Importance | Percentage | Rank |
| 1 | idiovol | 252020.6634 | 0.9522 | 0.3287 | 1.3333 |
| 2 | baspread | 111399.6761 | 0.5508 | 0.1489 | 1.9167 |
| 3 | retvol | 98918.6611 | 0.4642 | 0.1321 | 2.9167 |
| 4 | maxret | 19223.0792 | 0.1236 | 0.0234 | 6.4167 |
| 5 | mom6m | 14410.1369 | 0.0873 | 0.0183 | 6.5833 |
| 6 | label10 | 16302.6913 | 0.1241 | 0.0221 | 6.75 |
| 7 | mom12m | 15108.8604 | 0.0783 | 0.0193 | 7 |
| 8 | sich2 | 25173.3684 | 0.1675 | 0.039 | 7.25 |
| 9 | mom1m | 15667.66 | 0.0752 | 0.0196 | 7.5 |
| 10 | ep | 9464.183 | 0.0616 | 0.0119 | 11 |
| 11 | roaq | 6032.3019 | 0.0377 | 0.0078 | 14.9091 |
| 12 | dy | 8893.5185 | 0.055 | 0.0105 | 16.0833 |
| 13 | betasq | 9981.4297 | 0.0667 | 0.0118 | 19.4545 |
| 14 | beta | 6456.432 | 0.0424 | 0.0078 | 21.6364 |
| 15 | turn | 3273.5324 | 0.0167 | 0.0037 | 21.6667 |
| 16 | zerotrade | 2721.8782 | 0.0105 | 0.003 | 23.0833 |
| 17 | dolvol | 5352.8563 | 0.0335 | 0.0065 | 25.9091 |
| 18 | ill | 4037.6283 | 0.0224 | 0.0046 | 27.6364 |
| 19 | nanalyst | 2547.2677 | 0.0147 | 0.0031 | 28.0909 |
| 20 | label10.0 | 2044.8495 | 0.004 | 0.0017 | 32.25 |
| 21 | ill_hi_minus_low | 2156.603 | 0.0074 | 0.0025 | 32.4167 |
| 22 | nonborres | 1580.9666 | 0.0059 | 0.0019 | 32.75 |
| 23 | roeq | 5354.326 | 0.0376 | 0.0067 | 35.6364 |
| 24 | std_turn | 1618.6375 | 0.0077 | 0.0018 | 37.5 |
| 25 | dp | 1464.5543 | 0.006 | 0.0017 | 38.1667 |
| 26 | sich2_ret | 1707.6039 | 0.0057 | 0.0018 | 40.25 |
| 27 | bm.x | 1923.647 | 0.0103 | 0.0022 | 41.0909 |
| 28 | fgr5yr_hi_minus_low | 1434.709 | 0.0053 | 0.0016 | 41.8333 |
| 29 | fgr5yr | 1980.004 | 0.012 | 0.0025 | 42.6667 |
| 30 | ddurrg3m086sbea | 1412.1605 | 0.0048 | 0.0016 | 45.6667 |
| 31 | roavol | 8958.6767 | 0.0642 | 0.0111 | 47 |
| 32 | agr | 1188.3276 | 0.0048 | 0.0013 | 49.0909 |
| 33 | ndmanemp | 1114.8946 | 0.0044 | 0.0014 | 49.2727 |
| 34 | chcsho | 1119.6585 | 0.0046 | 0.0014 | 50.4545 |
| 35 | chmom | 1746.2888 | 0.009 | 0.002 | 51.4545 |
| 36 | turn_hi_minus_low | 1065.2049 | 0.0037 | 0.0012 | 55.5 |
| 37 | roic | 4237.5983 | 0.0303 | 0.0053 | 59.4545 |
| 38 | bm.y | 1311.4077 | 0.0055 | 0.0015 | 60.7273 |
| 39 | cusr0000sas | 997.8378 | 0.0047 | 0.0011 | 61.8333 |
| 40 | bm_hi_minus_low | 980.8402 | 0.0036 | 0.0011 | 64.8182 |
| 41 | chcsho_hi_minus_low | 844.7581 | 0.0031 | 9.00E-04 | 65.75 |
| 42 | age | 7391.9657 | 0.0557 | 0.0094 | 68.0909 |
| 43 | invest.x | 978.3864 | 0.004 | 0.001 | 68.2727 |
| 44 | ms_hi_minus_low | 903.585 | 0.0034 | 0.001 | 68.3636 |
| 45 | pchgm_pchsale_hi_minus_low | 841.0044 | 0.0032 | 9.00E-04 | 70.1667 |
| 46 | realestate_hi_minus_low | 672.6382 | 0.0029 | 8.00E-04 | 70.5833 |
| 47 | nanalyst_hi_minus_low | 901.2761 | 0.0037 | 0.0011 | 70.8182 |
| 48 | dolvol_hi_minus_low | 950.7877 | 0.0034 | 0.001 | 76.4545 |
| 49 | sfe_hi_minus_low | 969.5795 | 0.0034 | 0.0011 | 76.6364 |
| 50 | exszusx | 1110.7301 | 0.0042 | 0.0013 | 76.9091 |

# Appendix

**Table A.1 Firm Characteristics**

This table presents characteristics we reconstructed based on Green, Hand and Zhang (2017).

| Name | Description | Min | Max | Mean | Median | Skewness | Kurtosis |
|---|---|---:|---:|---:|---:|---:|---:|
| acc | Working capital accruals | -1.02 | 0.58 | -0.02 | -0.02 | -0.88 | 4.81 |
| aeavol | Abnormal earnings announcement volume | -1.00 | 21.69 | 0.87 | 0.30 | 3.64 | 18.51 |
| age | # years since first Compustat coverage | 1.00 | 56.00 | 12.72 | 9.00 | 1.36 | 1.55 |
| agr | Asset growth | -0.68 | 5.85 | 0.15 | 0.08 | 4.26 | 30.02 |
| baspread | Bid-ask spread | 0.00 | 0.91 | 0.05 | 0.03 | 5.15 | 38.11 |
| beta | Beta | -0.74 | 3.94 | 1.08 | 1.01 | 0.69 | 0.69 |
| bm | Book-to-market | -2.35 | 7.81 | 0.77 | 0.60 | 2.48 | 11.46 |
| cash | Cash holdings | 0.00 | 0.98 | 0.16 | 0.07 | 1.89 | 3.15 |
| cashdebt | Cash flow to debt | -7.71 | 2.23 | 0.07 | 0.13 | -4.14 | 25.99 |
| cashpr | Cash productivity | -520.62 | 600.28 | -1.90 | -0.73 | 0.89 | 29.15 |
| cfp | Cash flow to price ratio | -513.56 | 156.76 | 0.05 | 0.05 | -172.20 | 57390.53 |
| cfp_ia | Industry-adjusted cash flow to price ratio | -449.37 | 7031.61 | 13.09 | 0.00 | 21.92 | 479.82 |
| chadv | Change in dividend | -1.59 | 2.02 | 0.05 | 0.03 | 0.50 | 8.14 |
| chatoia | Industry-adjusted change in asset turnover | -1.43 | 1.19 | 0.00 | 0.00 | -0.15 | 4.74 |
| chcsho | Change in shares outstanding | -0.89 | 2.57 | 0.11 | 0.01 | 3.28 | 13.85 |
| chfeps | Change in forecasted EPS | -6.48 | 8.25 | 0.00 | 0.00 | 1.29 | 121.37 |
| chinv | Change in inventory | -0.29 | 0.37 | 0.01 | 0.00 | 1.10 | 6.77 |
| chnanalyst | Change in number of analysts | -12.00 | 9.00 | -0.01 | 0.00 | -0.60 | 9.46 |
| chtx | Change in tax expense | -0.12 | 0.16 | 0.00 | 0.00 | 0.35 | 13.06 |
| cinvest | Corporate investment | -26.83 | 27.87 | -0.02 | 0.00 | -2.17 | 244.24 |
| currat | Current ratio | 0.16 | 60.34 | 3.16 | 2.00 | 5.58 | 40.35 |
| depr | Depreciation/PP&E | 0.01 | 5.51 | 0.26 | 0.15 | 5.92 | 49.70 |
| disp | Dispersion in forecasted EPS | 0.00 | 10.00 | 0.15 | 0.04 | 6.48 | 58.18 |
| dy | Dividend to price | 0.00 | 0.35 | 0.02 | 0.00 | 2.67 | 10.86 |
| ear | Earnings announcement return | -0.46 | 0.51 | 0.00 | 0.00 | 0.26 | 3.17 |
| egr | Growth in common shareholder equity | -3.54 | 8.19 | 0.14 | 0.08 | 3.32 | 28.51 |
| ep | Earnings to price | -7.66 | 0.68 | -0.01 | 0.05 | -8.11 | 107.23 |
| fgr5yr | Forecasted growth in 5-year EPS | -43.50 | 99.41 | 16.35 | 14.50 | 1.50 | 5.47 |
| gma | Gross profitability | -0.84 | 1.78 | 0.37 | 0.33 | 0.81 | 1.52 |
| grcapx | Growth in capital expenditures | -13.89 | 55.54 | 0.89 | 0.14 | 5.60 | 45.95 |
| grltnoa | Growth in long term net operating assets | -0.61 | 1.18 | 0.09 | 0.06 | 1.64 | 7.48 |
| herf | Industry sales concentration | 0.01 | 1.00 | 0.08 | 0.05 | 3.10 | 11.91 |
| hire | Employee growth rate | -0.74 | 4.00 | 0.09 | 0.02 | 3.81 | 24.97 |
| idiovol | Idiosyncratic return volatility | 0.01 | 0.26 | 0.06 | 0.06 | 1.47 | 2.70 |
| ill | Illiquidity | 0.00 | 0.00 | 0.00 | 0.00 | 14.63 | 355.90 |
| indmom | Industry momentum | -1.00 | 3.56 | 0.14 | 0.12 | 1.26 | 5.39 |
| invest | Capital expenditures and inventory | -0.52 | 2.21 | 0.08 | 0.04 | 2.51 | 12.80 |
| lev | Leverage | 0.00 | 77.75 | 2.28 | 0.69 | 5.47 | 43.92 |
| Meanrec | Mean number of analysts | 1.00 | 4.50 | 2.22 | 2.20 | -0.01 | -0.32 |
| mom12m | 12-month momentum | -1.00 | 11.60 | 0.13 | 0.06 | 2.89 | 21.73 |
| mom1m | 1-month momentum | -0.70 | 2.11 | 0.01 | 0.00 | 1.16 | 7.77 |
| mom36m | 36-month momentum | -0.98 | 16.20 | 0.33 | 0.16 | 3.08 | 20.08 |
| ms | Financial statement score | 0.00 | 8.00 | 3.73 | 4.00 | -0.03 | -0.72 |
| mve | Size | 6.02 | 18.90 | 11.77 | 11.63 | 0.29 | -0.31 |
| mve_ia | Industry-adjusted size | -16608.51 | 133635.00 | -158.25 | -359.12 | 9.17 | 120.31 |

**Table A.1 (Continues)**

| Name | Description | Min | Max | Mean | Median | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| nanalyst | Number of analysts covering stock | 0.00 | 34.00 | 5.17 | 3.00 | 1.75 | 2.80 |
| nincr | Number of earnings increases | 0.00 | 8.00 | 1.00 | 1.00 | 2.15 | 6.35 |
| orgcap | Organizational capital | 0.00 | 0.18 | 0.01 | 0.01 | 2.73 | 11.60 |
| pchcapx_ia | Industry adjusted % change in capital expenditures | -237.42 | 1640.09 | 6.50 | -0.35 | 15.17 | 273.34 |
| pchcurrat | % change in current ration | -0.89 | 6.72 | 0.06 | -0.01 | 3.82 | 23.69 |
| pchdepr | % change in depreciation | -0.85 | 7.37 | 0.10 | 0.03 | 4.56 | 36.69 |
| pchgm_pchsale | % change in gross margin - % change in sales | -12.26 | 4.77 | -0.06 | 0.00 | -5.49 | 54.90 |
| pchsale_pchinvt | % change in sales - % change in inventory | -11.61 | 3.02 | -0.06 | 0.01 | -5.85 | 57.26 |
| pchsale_pchrect | % changes in sales - % change in A/R | -7.93 | 3.11 | -0.04 | 0.00 | -2.90 | 24.02 |
| pchsale_pchxsga | % change in sales - % change in SG&A | -3.50 | 4.34 | 0.02 | 0.00 | 3.42 | 31.72 |
| pctacc | Percent accruals | -64.75 | 71.43 | -0.65 | -0.27 | -1.90 | 34.80 |
| pricedelay | Price delay | -15.85 | 15.52 | 0.14 | 0.06 | 0.09 | 39.90 |
| ps | Financial statement score | 0.00 | 8.00 | 4.18 | 4.00 | 0.03 | -0.55 |
| rd_mve | R&D to market capitalization | 0.00 | 2.23 | 0.06 | 0.03 | 5.12 | 48.03 |
| rd_sale | R&D to sales | 0.00 | 283.48 | 0.61 | 0.03 | 23.58 | 733.49 |
| retvol | Return volatility | 0.00 | 0.27 | 0.03 | 0.02 | 2.42 | 8.82 |
| roaq | Return on assets | -0.48 | 0.16 | 0.00 | 0.01 | -3.40 | 16.22 |
| roavol | Earnings volatility | 0.00 | 0.85 | 0.03 | 0.01 | 5.31 | 41.10 |
| roe | Return on equity | -7.05 | 8.80 | 0.03 | 0.10 | -2.56 | 35.80 |
| roeq | Quarterly return on equity | -2.22 | 1.66 | 0.00 | 0.02 | -2.80 | 29.85 |
| roic | Return on invested capital | -21.24 | 1.01 | -0.08 | 0.07 | -10.40 | 149.98 |
| rsup | Revenue surprise | -4.51 | 2.33 | 0.02 | 0.01 | -3.83 | 64.41 |
| salecash | Sales to cash | 0.00 | 2503.48 | 52.60 | 10.60 | 7.68 | 73.99 |
| saleinv | Sales to inventory | 0.29 | 1031.22 | 25.91 | 7.59 | 6.96 | 62.95 |
| salerec | Sales to receivables | 0.00 | 594.00 | 11.68 | 5.94 | 5.25 | 31.16 |
| sfe | Scaled earnings forecast | -36.23 | 1.09 | -0.06 | 0.04 | -14.77 | 296.41 |
| sgr | Sales growth | -0.91 | 8.50 | 0.18 | 0.09 | 5.68 | 49.16 |
| sp | Sales to price | 0.00 | 54.59 | 2.32 | 1.10 | 4.51 | 29.91 |
| spi | Industry-adjusted sales to price | -0.66 | 0.19 | -0.01 | 0.00 | -5.20 | 40.89 |
| std_dolvol | Volatility of liquidity (dollar trading volume) | 0.18 | 2.74 | 0.86 | 0.79 | 0.75 | 0.18 |
| std_turn | Volatility of liquidity (share turnover) | 0.02 | 184.01 | 3.90 | 1.90 | 6.07 | 64.87 |
| stdcf | Cash flow volatility | 0.00 | 1882.88 | 9.88 | 0.14 | 11.94 | 178.00 |
| sue | Unexpected quarterly earnings | -5.20 | 1.70 | 0.00 | 0.00 | -13.43 | 550.18 |
| tang | Debt capacity/firm tangibility | 0.04 | 0.98 | 0.54 | 0.55 | -0.14 | 0.99 |
| tb | Tax income to book income | -27.70 | 15.36 | -0.10 | -0.03 | -4.47 | 66.74 |
| turn | Share turnover | 0.00 | 195.94 | 1.02 | 0.52 | 20.43 | 1384.13 |
| zerotrade | Zero trading days | 0.00 | 19.95 | 1.31 | 0.00 | 3.03 | 9.22 |

**Table A.2 Macroeconomic Variables**

Table A.2 demonstrates the macroeconomic variables we collect from McCracken's page on the website of Federal Reserve Bank of St. Louis. The variables are transformed following the recommended transformation methods. We excluded New Orders for Consumer Goods (ACOGNO), New Orders for Nondefense Capital Goods (ANDENOx) and Trade Weighted USD Index (TWEXMMTH) for data availability issues. In the end, we have 125 macroeconomic variables for our sample period from 196301:201912.

| Name | Description | Min | Max | Mean | Median | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| AAA | Moody's Seasoned Aaa Corporate Bond Yield | -1.18 | 1.29 | 0.00 | 0.00 | -0.26 | 5.52 |
| AAAFFM | Moody's Aaa Corporate Bond Minus FEDFUNDS | -6.27 | 5.73 | 2.07 | 2.21 | -0.85 | 1.17 |
| AMDMNOx | New Orders for Durable Goods | -0.20 | 0.21 | 0.00 | 0.00 | -0.14 | 3.19 |
| AMDMUOx | Unfilled Orders for Durable Goods | -0.03 | 0.05 | 0.00 | 0.00 | 0.53 | 1.07 |
| AWHMAN | Avg Weekly Hours : Manufacturing | 37.30 | 42.40 | 40.81 | 40.80 | -0.41 | 0.31 |
| AWOTMAN | Avg Weekly Overtime Hours : Manufacturing | -0.90 | 0.80 | 0.00 | 0.00 | -0.07 | 6.76 |
| BAA | Moody's Seasoned Baa Corporate Bond Yield | -1.02 | 1.57 | 0.00 | 0.00 | 0.55 | 6.78 |
| BAAFFM | Moody's Baa Corporate Bond Minus FEDFUNDS | -4.05 | 8.82 | 3.10 | 3.18 | -0.49 | 0.31 |
| BOGMBASE | Monetary Base | 0.00 | 0.06 | 0.00 | 0.00 | 15.95 | 271.36 |
| BUSINVx | Total Business Inventories | -0.02 | 0.07 | 0.00 | 0.00 | 1.65 | 22.45 |
| BUSLOANS | Commercial and Industrial Loans | 0.00 | 0.00 | 0.00 | 0.00 | 4.21 | 26.46 |
| CE16OV | Civilian Employment | -0.01 | 0.02 | 0.00 | 0.00 | -0.08 | 1.85 |
| CES0600000007 | Avg Weekly Hours : Goods-Producing | 37.20 | 41.80 | 40.32 | 40.30 | -0.28 | 0.28 |
| CES0600000008 | Avg Hourly Earnings : Goods-Producing | 0.00 | 0.00 | 0.00 | 0.00 | 5.03 | 44.16 |
| CES1021000001 | All Employees: Mining and Logging: Mining | -0.19 | 0.20 | 0.00 | 0.00 | 0.42 | 54.13 |
| CES2000000008 | Avg Hourly Earnings : Construction | 0.00 | 0.00 | 0.00 | 0.00 | 5.99 | 56.58 |
| CES3000000008 | Avg Hourly Earnings : Manufacturing | 0.00 | 0.00 | 0.00 | 0.00 | 5.08 | 36.95 |
| CLAIMSx | Initial Claims | -0.22 | 0.20 | 0.00 | 0.00 | 0.30 | 2.18 |
| CLF16OV | Civilian Labor Force | -0.01 | 0.01 | 0.00 | 0.00 | 0.11 | 2.62 |

**Table A.2 (Continues)**

| Name | Description | Min | Max | Mean | Median | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| CMRMTSPLx | Real Manu. and Trade Industries Sales | -0.03 | 0.05 | 0.00 | 0.00 | 0.01 | 1.22 |
| COMPAPFFx | 3-Month Commercial Paper Minus FEDFUNDS | -2.78 | 2.22 | 0.07 | 0.08 | -1.18 | 8.33 |
| CONSPI | Nonrevolving Consumer Credit to Personal Income | -0.01 | 0.01 | 0.00 | 0.00 | -0.34 | 23.21 |
| CP3Mx | 3-Month AA Financial Commercial Paper Rate | -6.29 | 3.03 | 0.00 | 0.00 | -2.31 | 42.14 |
| CPIAPPSL | CPI : Apparel | 0.00 | 0.00 | 0.00 | 0.00 | 4.33 | 24.18 |
| CPIAUCSL | CPI : All Items | 0.00 | 0.00 | 0.00 | 0.00 | 4.01 | 23.42 |
| CPIMEDSL | CPI : Medical Care | 0.00 | 0.00 | 0.00 | 0.00 | 3.01 | 15.77 |
| CPITRNSL | CPI : Transportation | 0.00 | 0.01 | 0.00 | 0.00 | 17.02 | 358.14 |
| CPIULFSL | CPI : All Items Less Food | 0.00 | 0.00 | 0.00 | 0.00 | 4.66 | 34.63 |
| CUMFNS | Capacity Utilization: Manufacturing | -3.77 | 2.22 | -0.01 | 0.03 | -0.82 | 3.76 |
| CUSR0000SA0L2 | CPI : All items less shelter | 0.00 | 0.00 | 0.00 | 0.00 | 7.95 | 98.51 |
| CUSR0000SA0L5 | CPI : All items less medical care | 0.00 | 0.00 | 0.00 | 0.00 | 4.24 | 26.39 |
| CUSR0000SAC | CPI : Commodities | 0.00 | 0.00 | 0.00 | 0.00 | 13.01 | 233.77 |
| CUSR0000SAD | CPI : Durables | 0.00 | 0.00 | 0.00 | 0.00 | 3.78 | 16.15 |
| CUSR0000SAS | CPI : Services | 0.00 | 0.00 | 0.00 | 0.00 | 3.74 | 17.25 |
| DDURRG3M086SBEA | Personal Cons. Exp: Durable goods | 0.00 | 0.00 | 0.00 | 0.00 | 4.73 | 31.52 |
| DMANEMP | All Employees: Durable goods | -0.05 | 0.03 | 0.00 | 0.00 | -1.43 | 9.48 |
| DNDGRG3M086SBEA | Personal Cons. Exp: Nondurable goods | 0.00 | 0.00 | 0.00 | 0.00 | 12.96 | 235.86 |
| DPCERA3M086SBEA | Real personal consumption expenditures | -0.03 | 0.02 | 0.00 | 0.00 | -0.12 | 2.94 |
| DSERRG3M086SBEA | Personal Cons. Exp: Services | 0.00 | 0.00 | 0.00 | 0.00 | 2.37 | 6.50 |
| DTCOLNVHFNM | Consumer Motor Vehicle Loans Outstanding | 0.00 | 0.03 | 0.00 | 0.00 | 9.94 | 121.28 |
| DTCTHFNM | Total Consumer Loans and Leases Outstanding | 0.00 | 0.06 | 0.00 | 0.00 | 19.57 | 405.68 |
| EXCAUSx | Canada / U.S. Foreign Exchange Rate | -0.06 | 0.11 | 0.00 | 0.00 | 0.52 | 8.22 |

**Table A.2 (Continues)**

| Name | Description | Min | Max | Mean | Median | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| EXJPUSx | Japan / U.S. Foreign Exchange Rate | -0.11 | 0.08 | 0.00 | 0.00 | -0.51 | 1.80 |
| EXSZUSx | Switzerland / U.S. Foreign Exchange Rate | -0.09 | 0.12 | 0.00 | 0.00 | -0.13 | 1.60 |
| EXUSUKx | U.S. / U.K. Foreign Exchange Rate | -0.11 | 0.10 | 0.00 | 0.00 | -0.46 | 2.81 |
| FEDFUNDS | Effective Federal Funds Rate | -6.63 | 3.06 | 0.00 | 0.01 | -2.39 | 47.47 |
| GS1 | 1-Year Treasury Rate | -3.91 | 1.90 | 0.00 | 0.00 | -1.51 | 18.10 |
| GS10 | 10-Year Treasury Rate | -1.76 | 1.61 | 0.00 | 0.00 | -0.43 | 5.94 |
| GS5 | 5-Year Treasury Rate | -2.03 | 1.86 | 0.00 | 0.00 | -0.41 | 6.68 |
| HOUST | Housing Starts: Total New Privately Owned | 6.17 | 7.82 | 7.22 | 7.29 | -0.98 | 1.00 |
| HOUSTMW | Housing Starts, Midwest | 4.08 | 6.38 | 5.55 | 5.66 | -0.89 | 0.29 |
| HOUSTNE | Housing Starts, Northeast | 3.58 | 5.98 | 5.04 | 5.04 | -0.34 | -0.11 |
| HOUSTS | Housing Starts, South | 5.44 | 7.08 | 6.42 | 6.44 | -0.57 | 0.36 |
| HOUSTW | Housing Starts, West | 4.37 | 6.47 | 5.78 | 5.84 | -0.93 | 0.62 |
| HWI | Help-Wanted Index for United States | -633.00 | 880.00 | 6.94 | 6.00 | 0.07 | 1.93 |
| HWIURATIO | Ratio of Help Wanted/No. Unemployed | -0.17 | 0.11 | 0.00 | 0.00 | -0.42 | 1.70 |
| INDPRO | IP Index | -0.04 | 0.03 | 0.00 | 0.00 | -0.99 | 5.11 |
| INVEST | Securities in Bank Credit at All Commercial Banks | 0.00 | 0.00 | 0.00 | 0.00 | 5.73 | 57.39 |
| IPB51222S | IP: Residential Utilities | -0.13 | 0.14 | 0.00 | 0.00 | -0.22 | 1.54 |
| IPBUSEQ | IP: Business Equipment | -0.08 | 0.05 | 0.00 | 0.00 | -1.01 | 5.36 |
| IPCONGD | IP: Consumer Goods | -0.03 | 0.04 | 0.00 | 0.00 | -0.02 | 1.85 |
| IPDCONGD | IP: Durable Consumer Goods | -0.11 | 0.13 | 0.00 | 0.00 | 0.02 | 6.60 |
| IPDMAT | IP: Durable Materials | -0.06 | 0.05 | 0.00 | 0.00 | -0.82 | 3.84 |
| IPFINAL | IP: Final Products (Market Group) | -0.03 | 0.03 | 0.00 | 0.00 | -0.32 | 1.90 |
| IPFPNSS | IP: Final Products and Nonindustrial Supplies | -0.03 | 0.03 | 0.00 | 0.00 | -0.48 | 2.31 |
| IPFUELS | IP: Fuels | -0.10 | 0.15 | 0.00 | 0.00 | 0.65 | 7.12 |
| IPMANSICS | IP: Manufacturing (SIC) | -0.05 | 0.03 | 0.00 | 0.00 | -0.91 | 4.25 |
| IPMAT | IP: Materials | -0.07 | 0.03 | 0.00 | 0.00 | -1.29 | 7.60 |

**Table A.2 (Continues)**

| Name | Description | Min | Max | Mean | Median | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| IPNCONGD | IP: Nondurable Consumer Goods | -0.02 | 0.02 | 0.00 | 0.00 | -0.08 | 0.38 |
| IPNMAT | IP: Nondurable Materials | -0.08 | 0.05 | 0.00 | 0.00 | -1.40 | 11.05 |
| ISRATIOx | Total Business: Inventories to Sales Ratio | -0.06 | 0.11 | 0.00 | 0.00 | 0.46 | 3.86 |
| M1SL | M1 Money Stock | 0.00 | 0.00 | 0.00 | 0.00 | 10.33 | 128.42 |
| M2REAL | Real M2 Money Stock | -0.02 | 0.03 | 0.00 | 0.00 | 0.73 | 3.58 |
| M2SL | M2 Money Stock | 0.00 | 0.00 | 0.00 | 0.00 | 5.65 | 50.69 |
| MANEMP | All Employees: Manufacturing | -0.03 | 0.02 | 0.00 | 0.00 | -1.47 | 6.60 |
| MZMSL | MZM Money Stock | 0.00 | 0.01 | 0.00 | 0.00 | 19.28 | 420.20 |
| NDMANEMP | All Employees: Nondurable goods | -0.02 | 0.01 | 0.00 | 0.00 | -1.12 | 4.60 |
| NONBORRES | Reserves Of Depository Institutions | -195.01 | 170.56 | -0.66 | -0.35 | -3.28 | 152.65 |
| NONREVSL | Total Nonrevolving Credit | 0.00 | 0.01 | 0.00 | 0.00 | 16.19 | 281.62 |
| OILPRICEx | Crude Oil, spliced WTI and Cushing | 0.00 | 0.73 | 0.01 | 0.00 | 19.83 | 456.21 |
| PAYEMS | All Employees: Total nonfarm | -0.01 | 0.01 | 0.00 | 0.00 | -0.48 | 2.68 |
| PCEPI | Personal Cons. Expend.: Chain Index | 0.00 | 0.00 | 0.00 | 0.00 | 3.09 | 11.28 |
| PERMIT | New Private Housing Permits (SAAR) | 6.24 | 7.79 | 7.18 | 7.21 | -0.75 | 0.43 |
| PERMITMW | New Private Housing Permits, Midwest (SAAR) | 4.34 | 6.23 | 5.50 | 5.60 | -0.78 | -0.02 |
| PERMITNE | New Private Housing Permits, Northeast (SAAR) | 4.06 | 6.05 | 5.07 | 5.08 | -0.23 | -0.26 |
| PERMITS | New Private Housing Permits, South (SAAR) | 5.37 | 7.01 | 6.32 | 6.36 | -0.30 | -0.43 |
| PERMITW | New Private Housing Permits, West (SAAR) | 4.57 | 6.63 | 5.79 | 5.85 | -0.89 | 0.57 |
| PPICMM | PPI: Metals and metal products: | 0.00 | 0.02 | 0.00 | 0.00 | 5.68 | 41.97 |
| REALLN | Real Estate Loans at All Commercial Banks | 0.00 | 0.00 | 0.00 | 0.00 | 8.25 | 106.81 |
| RETAILx | Retail and Food Services Sales | -0.07 | 0.06 | 0.00 | 0.00 | -0.30 | 4.93 |
| RPI | Real Personal Income | -0.05 | 0.04 | 0.00 | 0.00 | -0.64 | 22.59 |
| S_P__indust | S&P’s Common Stock Price Index: Industrials | -0.22 | 0.11 | 0.01 | 0.01 | -1.00 | 4.03 |

**Table A.2 (Continues)**

| Name | Description | Min | Max | Mean | Median | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| S_P_500 | S&P's Common Stock Price Index: Composite | -0.23 | 0.11 | 0.01 | 0.01 | -1.02 | 4.23 |
| S_P_div_yield | S&P's Composite Common Stock: Dividend Yield | -0.64 | 0.59 | 0.00 | -0.01 | 0.60 | 6.06 |
| S_P_PE_ratio | S&P's Composite Common Stock: Price-Earnings Ratio | -0.22 | 0.24 | 0.00 | 0.00 | 0.07 | 5.83 |
| SRVPRD | All Employees: Service-Providing Industries | -0.01 | 0.01 | 0.00 | 0.00 | 0.06 | 4.34 |
| T10YFFM | 10-Year Treasury C Minus FEDFUNDS | -6.51 | 3.85 | 1.04 | 1.21 | -1.11 | 2.12 |
| T1YFFM | 1-Year Treasury C Minus FEDFUNDS | -5.00 | 1.69 | 0.02 | 0.13 | -2.23 | 8.91 |
| T5YFFM | 5-Year Treasury C Minus FEDFUNDS | -6.31 | 3.16 | 0.70 | 0.87 | -1.35 | 3.27 |
| TB3MS | 3-Month Treasury Bill | -4.62 | 2.61 | 0.00 | 0.01 | -1.85 | 28.22 |
| TB3SMFFM | 3-Month Treasury C Minus FEDFUNDS | -5.37 | 0.68 | -0.49 | -0.25 | -2.56 | 9.28 |
| TB6MS | 6-Month Treasury Bill | -4.23 | 2.17 | 0.00 | 0.01 | -1.83 | 23.64 |
| | | | | | | | |
| TB6SMFFM | 6-Month Treasury C Minus FEDFUNDS | -5.01 | 1.19 | -0.35 | -0.13 | -2.57 | 10.00 |
| TOTRESNS | Total Reserves of Depository Institutions | 0.00 | 1.25 | 0.00 | 0.00 | 18.26 | 364.58 |
| UEMP15OV | Civilians Unemployed - 15 Weeks & Over | -0.18 | 0.24 | 0.00 | 0.00 | 0.37 | 1.27 |
| UEMP15T26 | Civilians Unemployed for 15-26 Weeks | -0.36 | 0.29 | 0.00 | 0.00 | -0.05 | 0.92 |
| UEMP27OV | Civilians Unemployed for 27 Weeks and Over | -0.21 | 0.28 | 0.00 | 0.00 | 0.29 | 1.21 |
| UEMP5TO14 | Civilians Unemployed for 5-14 Weeks | -0.22 | 0.23 | 0.00 | 0.00 | 0.28 | 1.18 |
| UEMPLT5 | Civilians Unemployed - Less Than 5 Weeks | -0.22 | 0.27 | 0.00 | 0.00 | -0.01 | 1.34 |
| UEMPMEAN | Average Duration of Unemployment (Weeks) | -2.70 | 2.50 | 0.01 | 0.00 | -0.08 | 2.25 |

**Table A.2 (Continues)**

| Name | Description | Min | Max | Mean | Median | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| UMCSENTx | Consumer Sentiment Index | -14.70 | 17.30 | 0.01 | 0.00 | 0.01 | 2.06 |
| UNRATE | Civilian Unemployment Rate | -0.70 | 0.90 | 0.00 | 0.00 | 0.52 | 1.98 |
| USCONS | All Employees: Construction | -0.04 | 0.06 | 0.00 | 0.00 | 0.20 | 5.92 |
| USFIRE | All Employees: Financial Activities | -0.01 | 0.01 | 0.00 | 0.00 | -0.51 | 1.20 |
| USGOOD | All Employees: Goods-Producing Industries | -0.02 | 0.02 | 0.00 | 0.00 | -1.26 | 4.61 |
| USGOVT | All Employees: Government | -0.01 | 0.02 | 0.00 | 0.00 | 0.67 | 8.38 |
| USTPU | All Employees: Trade, Transportation & Utilities | -0.01 | 0.01 | 0.00 | 0.00 | -0.43 | 1.33 |
| USTRADE | All Employees: Retail Trade | -0.01 | 0.01 | 0.00 | 0.00 | -0.24 | 2.71 |
| USWTRADE | All Employees: Wholesale Trade | -0.01 | 0.01 | 0.00 | 0.00 | -0.57 | 0.94 |
| VXOCLSx | VXO | 8.02 | 67.15 | 19.05 | 17.48 | 1.96 | 6.97 |
| W875RX1 | Real personal income ex transfer receipts | -0.06 | 0.04 | 0.00 | 0.00 | -1.77 | 32.51 |
| WPSFD49207 | Producer Price Index by Commodity: Final Demand: Finished Goods | 0.00 | 0.00 | 0.00 | 0.00 | 6.62 | 61.76 |
| WPSFD49502 | PPI: Final Demand: Personal Consumption Goods (Finished Consumer Goods) | 0.00 | 0.00 | 0.00 | 0.00 | 7.10 | 70.09 |
| WPSID61 | PPI: Intermediate Demand by Commodity Type: Processed Goods for Intermediate Demand | 0.00 | 0.00 | 0.00 | 0.00 | 7.84 | 77.81 |
| WPSID62 | PPI: Intermediate Demand by Commodity Type: Unprocessed Goods for Intermediate Demand | 0.00 | 0.04 | 0.00 | 0.00 | 7.48 | 69.10 |

**Table A.3 Probability Adjusted Value Weight Long-Short Portfolio Excluding Bottom 50% ME**

This table presents the economic performance of the probability adjusted allocation for the top 50% capitalization stocks. On average, the pool of stocks reduces from 4880 stocks to 2440 stocks from 1963 to 2019. The allocation rule is the same as described in Table 15.

| Model | Mean | SD | Skewness | SR | CEQ | Cumulative Return | Min | Max DD |
|---|---|---|---|---|---|---|---|---|
| Buy-Hold | 0.01 | 0.04 | -0.55 | 0.12 | 0.00 | 1989% | -0.23 | 0.54 |
| ANN1 128 | 0.01 | 0.05 | 0.03 | 0.22 | 0.01 | 46733% | -0.23 | 0.59 |
| ANN1 16 | 0.02 | 0.04 | -0.20 | 0.37 | 0.02 | 3592625% | -0.21 | 0.35 |
| ANN1 32 | 0.02 | 0.05 | 0.18 | 0.29 | 0.01 | 1600039% | -0.24 | 0.56 |
| ANN1 64 | 0.01 | 0.05 | -0.18 | 0.28 | 0.01 | 354674% | -0.20 | 0.74 |
| ANN2 128 | 0.00 | 0.08 | 0.10 | 0.05 | 0.00 | 84% | -0.27 | 0.82 |
| ANN2 32 | 0.02 | 0.06 | 0.29 | 0.30 | 0.02 | 5141824% | -0.31 | 0.49 |
| ANN2 64 | 0.01 | 0.08 | -0.29 | 0.11 | 0.01 | 3677% | -0.41 | 0.95 |
| ANN3 128 | 0.02 | 0.05 | 0.26 | 0.34 | 0.01 | 16885% | -0.14 | 0.34 |
| ANN3 64 | 0.02 | 0.05 | 0.08 | 0.31 | 0.01 | 1434913% | -0.20 | 0.40 |
| ANN4 128 | 0.02 | 0.07 | 0.04 | 0.33 | 0.02 | 74908663% | -0.30 | 0.47 |
| DART2 100 | 0.01 | 0.05 | -0.14 | 0.11 | 0.00 | 1716% | -0.25 | 0.85 |
| DART4 100 | 0.01 | 0.05 | 0.14 | 0.20 | 0.01 | 35532% | -0.22 | 0.63 |
| DART6 100 | 0.02 | 0.05 | -0.22 | 0.34 | 0.02 | 2973936% | -0.22 | 0.61 |
| DART8 100 | 0.02 | 0.05 | -0.13 | 0.30 | 0.01 | 2402339% | -0.24 | 0.57 |
| DRF2 200 | 0.01 | 0.05 | 0.08 | 0.15 | 0.01 | 4460% | -0.17 | 0.44 |
| DRF4 200 | 0.01 | 0.06 | 0.03 | 0.11 | 0.00 | 2114% | -0.30 | 0.86 |
| DRF6 200 | 0.01 | 0.06 | -0.14 | 0.20 | 0.01 | 58850% | -0.25 | 0.82 |
| DRF8 200 | 0.02 | 0.06 | -0.24 | 0.28 | 0.01 | 1738895% | -0.30 | 0.58 |
| GBM2 100 | 0.01 | 0.05 | -0.69 | 0.11 | 0.00 | 1985% | -0.39 | 0.86 |
| GBM4 100 | 0.01 | 0.05 | -0.27 | 0.24 | 0.01 | 186228% | -0.30 | 0.79 |
| GBM6 100 | 0.02 | 0.06 | -0.07 | 0.30 | 0.01 | 2293686% | -0.23 | 0.75 |
| GBM8 100 | 0.02 | 0.05 | -0.07 | 0.36 | 0.02 | 17832251% | -0.25 | 0.60 |