# The impact of Methodological choices on Machine Learning Portfolios

Vaibhav Lalwani* Vedprakash Meshram† Varun Jindal‡

This Version: May 21, 2024

**Preliminary and Incomplete.**
**Please do not cite or circulate.**

**Abstract**

We explore the impact of research design choices on the profitability of Machine learning investment strategies. Results from more than a thousand strategies show that considerable variation is induced by methodological choices on strategy returns. The non-standard errors of machine-learning strategies are often higher than the standard errors and remain sizeable even after controlling for some high-impact decisions. While eliminating micro-caps and using value-weighted portfolios reduces non-standard errors, their size is still quantitatively comparable to the traditional standard errors.

*XLRI Xavier School of Management Delhi-NCR vaibhav@xlri.ac.in
†Goa Institute of Management, Goa vedprakash@gim.ac.in
‡Indian Institute of Management Bangalore varun.jindal@iimb.ac.in

# 1 Introduction

Recent advancements in computational finance have underscored the potential of machine learning to enhance the predictive accuracy of return forecasting models. By leveraging non-linear relationships and high-dimensional data, these methods have shown promise in uncovering complex sources of predictability, hitherto overlooked by traditional models. Compared to the traditional characteristic-based portfolio sorting methods, machine learning tools allow researchers a convenient way to incorporate a large number of predictors in a predictive set-up. As the zoo of variables predicting returns is already very large, machine-learning methods are naturally adapted to handle the challenges of combining signals from multiple variables into a single all-encompassing strategy. Theoretical benefits aside, empirical results in Gu et al. (2020), Azevedo et al. (2023),Bali et al. (2023),Bianchi et al. (2021),Blitz et al. (2023), Cakici et al. (2023a), Cakici et al. (2023b) and Cakici et al. (2024) convincingly demonstrate that machine learning can be used to predict asset returns across various asset classes, geographies, and horizons.

Therefore, there is ample evidence that suggests that researchers can use ML tools to develop better return forecasting models. However, a researcher needs to make certain choices when using machine learning in return forecasting. These choices include, but are not limited to the size of training and validation windows, the outcome variable, data filtering, weighting, and the set of predictor variables. In a sample case with 10 decision variables, each offering two decision paths, the total specification are $2^{10}$, i.e. 1024. Accommodating more complex choices can lead to thousands of possible paths that the research design could take. While most studies integrate some level of robustness checks, keeping up with the entire universe of possibilities is virtually impossible. Further, with the computationally intensive nature of machine learning tasks, it is extremely challenging to explore the impact of all of these choices even if a researcher wishes to. Therefore, some of these calls are usually left to the better judgment of the researcher. While the sensitivity of findings to even apparently harmless empirical decisions is well-acknowledged in the literature[1], we have only very recently begun to acknowledge the size of the problem at hand. Menkveld et al. (2024) coin the term to *Non-standard errors* to denote the uncertainty in estimates due to different research choices. Studies like Soebhag et al. (2023) and Walter et al. (2023), and Fieberg et al. (2024) show that non-standard errors can be as large, if not larger

[1] A casual look at the robustness checks sections in contemporary finance papers should convince a reader of the veracity of this statement

than traditional standard errors. This phenomenon raises important questions about the reproducibility and reliability of financial research. It underscores the need for a possibly more systematic approach to the choice of methodological specifications and the importance of transparency in reporting research methodologies and results. As even seemingly innocuous choices can have a significant impact on the final results, unless we conduct a formal analysis of all (or at least, most) of the design choices together, it will be hard to know which choices matter and which do not through pure intuition.

Even in asset-pricing studies that use single characteristic sorting, there are thousands of possible choices (Walter et al. (2023) use as many as 69,120 potential specifications). Extending the analysis to machine learning-based portfolios, the possible list of choices (and their possible impact) further expands. Machine-learning users have to make many additional choices for modeling the relationship between returns and predictor characteristics. With the number of machine learning models available, (see Gu et al. (2020) for a subset of the possible models), it would not be unfair to say that scholars in the field are spoilt for choices. As argued by Harvey (2017) and Coqueret (2023), such a large number of choices might exacerbate the publication bias in favor of positive results.

The question that looms large is - what should we do about non-standard errors? In this paper, we take the approach suggested by Coqueret (2023) and Fabozzi and Prado (2018), among others, i.e. to consider the entire possible set of results under different decision paths. While Soebhag et al. (2023) suggests standardization of methodological choices, Walter et al. (2023) suggests reporting the entire distribution of outcomes. We believe, however, that exploring all (or even most) possible paths is not possible at all times. Even in portfolio sorts (Walter et al., 2023), the process of estimating all possible outcomes is very computationally intensive, albeit doable. However, in some cases, the computational burden may be too excessive, particularly for researchers without access to (relatively) costly infrastructure. In the case of machine learning-based portfolios, a back of the envelope suggests an overall computational time of 3-6 months on a single computer for even a modest number of decision variables. Unlike Coqueret (2023) or Walter et al. (2023), our objective in this study is not to suggest better ways of dealing with methodological variations. Instead, we aim to quantify the impact of various decisions on the final performance of portfolios created from machine learning forecasts. In doing so, we hope to shed light on the trade-offs involved in various methodological choices during the construction

of machine learning-based investment strategies.

Following Walter et al. (2023), we define non-standard errors as the inter-quantile range of the returns obtained from various methodological choices. To this regard, we evaluate the "non-standard errors" arising from eight decision criteria, resulting in 384 total paths for each machine learning model. Our results show a substantial variation in returns generated by machine learning portfolios. Depending on how the forecasts are generated, the non-standard errors of returns can be as large as 4.8 times the standard errors. Our findings show that non-standard errors are sizeable, and are often as large as the standard errors.

Our study contributes to three strands of the literature in financial economics. First, we contribute to the literature on empirical models of return forecasting. The typical return forecasting study involves developing models of return prediction and comparing their performance against benchmarks. Such studies typically focus on the benefits of methodological improvements in the forecasting model, but not so much on other choices such as sampling window and industry filtering. Our primary contribution is to provide a comprehensive overview of the impact of various methodological choices on the outcome, i.e. portfolio performance.

Interest in applications of Machine learning in Finance has grown substantially in the last decade or so. Since the seminal work of Gu et al. (2020), many variants of machine learning models have been used to predict asset returns. Our second contribution is to this growing body of literature. That there are many choices while using ML in return forecasting is well understood. But are the differences between specifications large enough to warrant caution? Avramov et al. (2023) shows that removing certain types of stocks considerably reduces the performance of machine-learning strategies. We expand this line of thought using a broader set of choices that include various considerations that hitherto researchers might have ignored. By providing a big-picture understanding of how the performance of machine learning strategies varies across decision paths, we conduct a kind of large-scale sensitivity analysis of the efficacy of machine learning in return forecasting. Additionally, by systematically analyzing the effects of various methodological choices, we can understand which factors are most influential in determining the success of a machine learning-based investment strategy.

Finally, we also add to the nascent literature in finance that explores choice-induced variation in research outcomes (i.e. non-standard errors). Research as far back as Leamer (1983) has tried to understand the role of choices in statistical out-

comes. However, the analysis of Leamer and Leonard (1983) and Sala-I-Martin (1997) deals with uncertainty around the choice of dependent variables. The overall choices available to a researcher relate to issues far beyond variable selection. Now that machine learning and other statistical tools that can handle large dimension datasets are available, at least in forecasting, other decisions could play a bigger role in inducing methodological variation vis-a-vis variable selection. Our analytical approach can efficiently accommodate such multi-choice settings.

To summarise, we find that the choices regarding the inclusion of micro-caps and penny stocks and the weighting of stocks have a significant impact on average returns. Further, an increase in sampling window length yields higher performance, but large windows are not needed for Boosting-based strategies. Based on our results, we argue that financials and utilities should not be excluded from the sample, at least not when using machine learning. Certain methodological choices can reduce the methodological variation around strategy returns, but the non-standard errors remain sizeable.

The rest of the paper proceeds as follows. Section 2 provides a background of the data and methods used. We then present our empirical results in Section 3. Finally, Section 4 concludes.

## 2 Data and Methodology

### 2.1 Data

Our dataset consists of the standard sample used in most asset-pricing studies. It consists of common stocks traded at the NYSE, AMEX, and NASDAQ stock exchanges. We collect the data on 200+ predictors used by Chen and Zimmermann (2021) from their website - opensourceassetpricing.com. Our returns are adjusted for delisting (Shumway, 1997), and following , we replace missing values of predictors with their monthly cross-sectional average. Our sample period is from 1957 to 2023. Like Walter et al. (2023), we do not consider the possible methodological variation in the construction of predictor variables and treat them as given. One could argue that even the choice of variables is expected to introduce methodological variation. As we are dealing with forecasting, that too with ML methods, we assume that a researcher would like to use the maximum possible set of variables available to them. So, we do not consider the possibility that a particular researcher would like to use a subset of the total variables available, although we do acknowledge that the choice of variables

can have a significant impact.

### 2.2 Methodology

We now describe the general version of the forecasting model used for return prediction, along with the methodological choices involved in estimating the parameters of the models.

Like Gu et al. (2020), our baseline return prediction model can be express in the form:

$$r_{i,t+1} = E_t(r_{i,t+1}) + \epsilon_{i,t+1} \tag{1}$$

where

$$E_t(r_{i,t+1}) = f(z_{i,t}) \tag{2}$$

where $i$ refers to the stock 'i' and $t$ refers to the months t=1,...T. Our objective is to estimate a function that minimises out-of-sample predictive errors for realized $r_{i,t+1}$ (Gu et al., 2020). As discussed, our basic dataset is the typical unbalanced panel of stocks used in asset-pricing studies. However, the actual data used in the estimation depend on the aforementioned methodological choices.

In the current study, we use eight choices that researchers have to consider in a study on return predictability. Some of these choices follow from typical asset-pricing studies like the ones discussed in Soebhag et al. (2023). Some are unique to return forecasting or Machine learning methods. Our decision choices are as follows:

1. **Training Window** An important decision that researchers have to make for a return forecasting model is the size and type of the model training window. Both rolling and expanding windows are popular. While Rasekhschaffe and Jones (2019) use a rolling window, Gu et al. (2020) use expanding windows. Further, with rolling windows, the ideal duration of the sampling window is unknown. So, researchers often use data in multiples of five years. In the current draft, we use three alternatives - 5, 10, and 15-year rolling windows. We plan on adding more complex sampling window schemes in the next version of the manuscript.

2. **Size Filter** We also take into account the decision to include or exclude micro-caps. From a practical standpoint, the exclusion of small stocks is highly relevant as such stocks are often illiquid, costly to trade, and even costlier to short.

While Kelly et al include such micro-caps in their base results, Avramov et al. (2023) show that exclusion of such stocks can significantly dent the performance of ML strategies. We impose a 20% NYSE breakpoint size filter in cases where micro-caps are excluded. We assume that this filter is applied before the model is trained.

3. **Price Filter** Harvey and Liu (2020) show that only about 18% of the studies impose a price filter to exclude small stocks. As discussed by Walter et al. (2023), price exclusions are quite different from size-based filtering as about 40% of stocks below $5 in their sample are not excluded in the size filter. Therefore, a price filter involves filtering out stocks below $5 in price.

4. **Utilities** Utilities firms are often excluded on account of the regulatory environment that they operate under. Many predictors in return forecasting use accounting data. Therefore, including firms that are expected to operate in different accounting environments can create problems. We do not have any ex-ante preference on the inclusion or exclusion of utilities. We believe it would be interesting to check if this filter has any significant impact on the returns of ML strategies.

5. **Financials** The case for the exclusion of financials is arguably stronger than that for Utilities. Financial firms are different from other firms in terms of their business model as well as the meaning of various accounting indicators. Financial firms are often highly leveraged and incomparable to firms in other industries. For studies that try to estimate premia related to underlying accounting characteristics, this can make a huge difference. However, in large-scale ML models, one could very well argue that the heterogeneity of financial stocks can be explicitly modelled by the non-linear interactions within the models. Therefore, there is a case to be made to include financials in the analysis.

6. **Age Filter** We also consider a filter where we only include firms that have at least two years of historical data. This filter is for controlling backfilling biases in the dataset (Banz and Breen, 1986) and is especially relevant for return forecasting as we only wish to consider information that was available to an investor while training the model.

7. **Weighting** Two common weighting procedures are used in most papers - equal and value-weighted. While value-weighted portfolios require less rebalancing

and load heavily on larger stocks, equally weighted portfolios are costlier to trade but are more diversified. As reported by Harvey and Liu (2020), about 40% of studies provide evidence using both equal and value-weighted portfolios. We include this filter to understand the extent to which the returns to ML strategies can vary with weighting.

8. **Quantiles** Paper on machine learning strategies typically considers decile portfolios, but quintiles are also popular in asset pricing. A smaller number of portfolios could reduce portfolio turnovers but each portfolio is likely to have a large number of assets, including the ones where the underlying signal is relatively weaker. Assuming a monotonous relationship between expected returns and ML forecasts, deciles are likely to yield better returns compared to quintiles, but it would be interesting to put this argument to the test.

These eight decisions imply 384 paths for each machine-learning model. The choice of machine learning estimator can also be considered a significantly important decision by the researcher. Barring a few exceptions, we do not consider the choice of model as a separate decision due to significant heterogeneity between the models. Instead, we choose to report the results of the 384 paths across different models. Our choice of estimators is motivated by existing research in asset pricing. Gu et al. (2020) and others show that methods that incorporate non-linear interactions among variables perform better than linear models. Therefore, we consider non-linear machine learning models in our study. In the current draft, we consider Random Forests, Gradient Boosting, and Neural Networks for training the return forecasting model.

In what follows, we provide a brief explanation of the models used in our study. We also discuss the values of hyperparameters that we consider in training the ML models.

1. **Random Forests** A random forest model is an ensemble method that aggregates from multiple decision trees (Breiman, 2001). It follows the process of bootstrapped aggregation, or Bagging. Bagging involves drawing multiple samples from the data, training individual decision tree models on these data, and then aggregating the forecasts generated from all the models. Like many other machine learning methods, users need to provide values of hyper-parameters. The choice of hyper-parameters can be considered a source of methodological variation. However,we do not consider it so because methodological variation should be introduced by decisions where multiple choices exist and each of

those choices are justifiable under some pretext. This condition, in our opinion, does not hold for hyperparameters. While there may not be much theoretical guidance available, researchers do not go on randomly choosing the values of parameters. They either use some validation process, such as a holdout validation (Gu et al. (2020)) or they will be guided by the values used in prior literature. For our study, we take guidance from the values used in Gu et al. (2020). We set tree depth as six and the number of trees as 300.

2. **Gradient Boosting** Gradient Boosting (or Boosting) is another ensemble method that combines the forecasts from multiple trees into a single average. However, the process is different from Bagging. Boosting builds an additive model by sequentially adding trees to a base model with low depth. At each following step, a shallow tree is fitted to the errors of the previous model. This procedure is continued till the final model is obtained (subject to hyperparameters). We use the XGBoost algorithm of Chen and Guestrin (2016) to train the models. We set the maximum tree depth as 6 and the number of trees (or rounds) as 300.

3. **Neural Networks** Neural Networks (or Artificial Neural Networks) are one of the most popular machine learning methods. Neural networks are the go to method for practical machine learning applications, including computer vision and Generative Artificial Intelligence tools like Large Language Models (e.g. Chat-GPT). Following studies like Gu et al. (2020) and (Avramov et al., 2023), we use feed-forward artificial neural networks with three hidden layers with 32, 16, and 8 neurons respectively. We use the ReLU activation function for all nodes. To reduce over-fitting, we use $l_1$ penalization and batch normalization for all trained models. To reduce the effect of random seed generation on forecasts, we follow Gu et al. (2020) and generate ten sets of forecasts with different seeds and average them out to arrive at a final ensemble forecast.

The methods and choices discussed above yield a common output, i.e. the monthly forecasts of returns for the next year. We report the average returns generated by strategies in given groups as a metric of portfolio performance. As we show later, we derive similar results even if we use the Sharpe ratio. We also report the standard and non-standard errors of the strategies used in our study. The standard error is the Newey and West (1987) standard errors of the time series average of individual portfolio returns, averaged across specifications. Following Menkveld et al. (2024)

and Walter et al. (2023), we define the non-standard errors as the inter-quantile range of average portfolio returns across all methodological specifications. We report the results of our analysis in the next section.

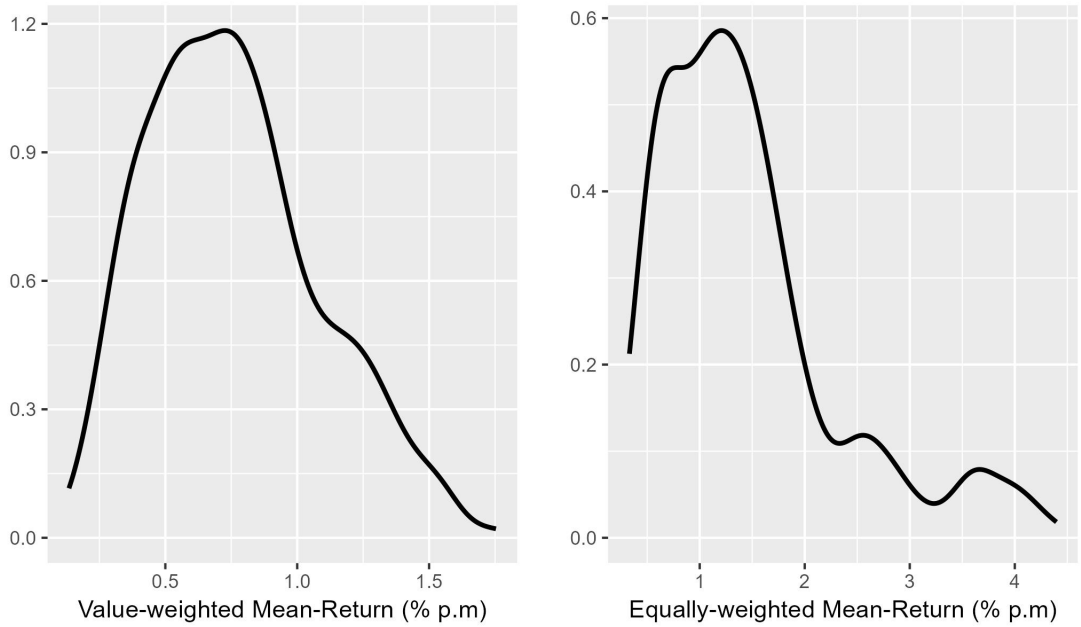
## 3 Results

Figure 1 shows the distribution of returns across various specifications. We observe a non-trivial variation in the monthly average returns observed across various choices. The variation appears to be much larger for equally-weighted portfolios compared to value-weighted portfolios, a result we find quite intuitive. The figure also points towards a few large outliers. It would be interesting to further analyze if these extreme values are driven by certain specification choices or are random. The variation in returns could be driven by the choice of the estimator. Studies like Gu et al. (2020) and Azevedo et al. (2023) report significant differences between returns from using different Machine Learning models. Therefore, we plot the return variation after separating models in Figure 2. Figure 2 makes it apparent that there is a considerable difference between the mean returns generated by different ML models. In our sample, Boosted Trees achieve the best out-of-sample performance, closely followed by Neural Networks. Random Forests appear to deliver much lower performance compared to the other two model types. Also, Figure 2 shows that the overall distribution of performance is similar for raw returns as well as Sharpe Ratios. Therefore, for the rest of our analysis, we consider long-short portfolio returns as the standard metric of portfolio performance.

All in all, there is a substantial variation in the returns generated by long-short machine learning portfolios. This variation is independent of the performance variation due to choice of model estimators. We now shift our focus toward understanding the impact of individual decisions on the average returns generated by each of the specifications. Therefore, we estimate the average of the mean returns for all specifications while keeping certain choices fixed. These results are in Table 1.

The results in Table 1 show that some choices impact the average returns more than others. Equal weighting of stocks in the sample increases the average returns. So does the inclusion of smaller stocks. The inclusion of financial and utilities appears to have a slightly positive impact on the overall portfolio Performance. Just like a size filter, the exclusion of low-price stocks tends to reduce overall returns. Further, grouping stocks in ten portfolios yields better performance compared to quintile sort-

Figure 1: The distribution of mean portfolio returns (% p.m) across all specifications

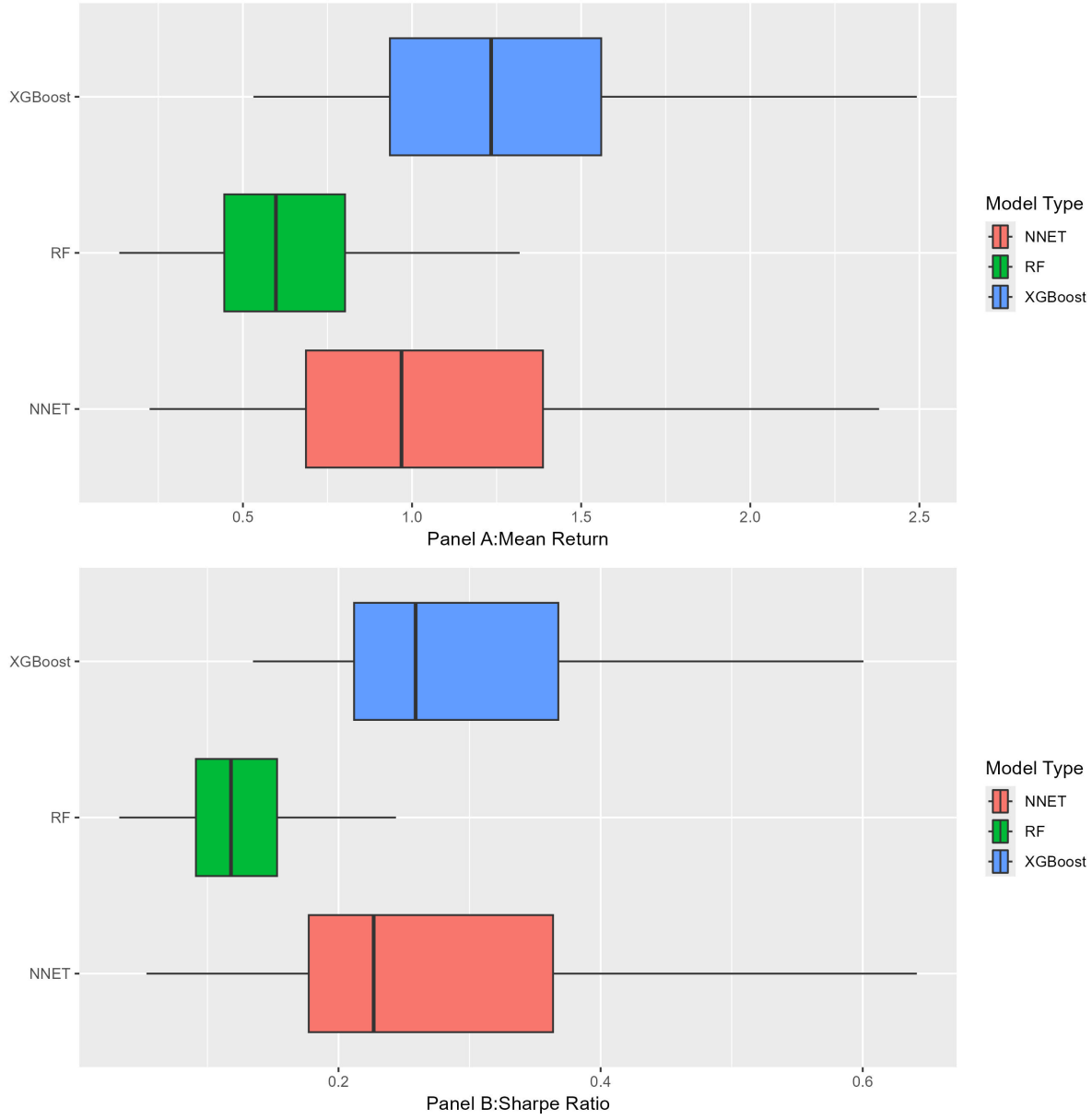



This figure plots the distribution of monthly average returns generated by equal and value-weighted Machine-Learning portfolios. The returns are in percentage(%) per month.

ing. On average, larger training windows appear to be better. However, this seems to be true largely for Neural Networks. For Neural Networks, the average return increases from 0.87% to 1.41% per month. For boosting, the gain is from 1.41% to 1.45%. XGBoost works well with just five years of data. It takes at least 15 years of data for Neural Networks to achieve the same performance. Interestingly, while Gu et al. (2020) and (Avramov et al., 2023) both use Neural Networks with a large expanding training window, our results show that similar performance can be achieved with a much smaller data set (but with XGBoost). Finally, the process of keeping only stocks with at least two years of data reduces the returns, but as discussed, this filter makes our results more applicable to real-time investors.

While controlling for choices individually yields important insights, there may be interactions at play. For example, the size and price filters have substantial overlaps. Therefore, we need to control for all other factors to verify the conditional impact of an individual choice. For this purpose, we run a multiple linear regression with the average return of each of our specifications as the dependent variable. We create dummy variables for each of the choices as the independent variables. In the Table 2, we report the OLS estimates of this regression.

Figure 2: Box-Plots of Model-wise Mean returns and Sharpe Ratios for all methodological choices




This figure reports the distribution of monthly average returns and Sharpe Ratios generated by Machine-Learning portfolios. The results are segregated according the Machine-learning model used in training the forecasting model. The returns are in percentage(%) per month.

Table 1: Average returns after keeping certain choices fixed

This table reports the summary of the monthly average returns generated by various Machine Learning Portfolio Strategies considered in our study. We report the average returns generated by keeping certain choices constant. The results contain combined as well as model-wise results.

| | | Average of Mean returns (% p.m.) | | | |
|---|---|---|---|---|---|
| **Decision Type** | **Decision Value** | **Overall** | **NNET** | **RF** | **XGBoost** |
| Weighting | Equal | 1.46 | 1.62 | 0.89 | 1.88 |
| | Value | 0.76 | 0.75 | 0.52 | 1.00 |
| Size Filter | No | 1.38 | 1.49 | 0.90 | 1.74 |
| | Yes | 0.84 | 0.88 | 0.51 | 1.14 |
| Financials Filter | No | 1.08 | 1.17 | 0.69 | 1.38 |
| | Yes | 1.14 | 1.20 | 0.72 | 1.49 |
| Utilities Filter | No | 1.10 | 1.20 | 0.69 | 1.40 |
| | Yes | 1.12 | 1.17 | 0.72 | 1.47 |
| Price Filter | No | 1.31 | 1.44 | 0.84 | 1.64 |
| | Yes | 0.91 | 0.93 | 0.57 | 1.23 |
| Age Filter | No | 1.17 | 1.28 | 0.73 | 1.50 |
| | Yes | 1.05 | 1.09 | 0.68 | 1.37 |
| Quantiles | 5 | 0.92 | 0.99 | 0.58 | 1.18 |
| | 10 | 1.30 | 1.38 | 0.83 | 1.69 |
| Training window size | 5 | 0.98 | 0.87 | 0.65 | 1.41 |
| | 10 | 1.14 | 1.29 | 0.70 | 1.44 |
| | 15 | 1.21 | 1.41 | 0.77 | 1.45 |

The coefficient estimates in Table 2 largely confirm what we observed earlier. The intercept refers to the average returns generated by a specification assuming the values of all dummy variables are zero. Therefore, for the results in the first column, the intercept value of 1.38 signifies that a model without size filter yields 1.38% return per month on average, and then the coefficient on the size filter shows that including the size filter reduces the (average of) average return by around 0.53%.

Therefore, based on the results in Table 2, the financial and utilities filter does not have a significant impact on the average mean returns. Applying both size and price filters reduces the returns by around 0.53% and 0.39% respectively. XGBoost models generate forecasts that yield higher portfolio returns compared to Neural Networks. Longer sampling windows also carry significant utility for model training.

Table 2: This table reports the results of the regression of average return on the dummy variables of various specification choices. Value in the parentheses contain the t-statistic of the coefficients

| | *Dependent variable:* | | |
|---|---|---|---|
| | Mean Return | | |
| | (1) | (2) | (3) |
| Intercept | $1.380^{***}$ | $1.600^{***}$ | $1.540^{***}$ |
| | (47.300) | (33.200) | (28.300) |
| Size Filter | $-0.533^{***}$ | $-0.533^{***}$ | $-0.533^{***}$ |
| | $(-13.000)$ | $(-13.600)$ | $(-15.500)$ |
| Fin Filter | | 0.058 | $0.058^{*}$ |
| | | (1.480) | (1.680) |
| Util Filter | | 0.024 | 0.024 |
| | | (0.615) | (0.700) |
| Price Filter | | $-0.395^{***}$ | $-0.395^{***}$ |
| | | $(-10.100)$ | $(-11.400)$ |
| Age Filter | | $-0.128^{***}$ | $-0.128^{***}$ |
| | | $(-3.270)$ | $(-3.720)$ |
| RF | | | $-0.482^{***}$ |
| | | | $(-11.400)$ |
| XGBOOST | | | $0.249^{***}$ |
| | | | (5.900) |
| Est 10 Years | | | $0.164^{***}$ |
| | | | (3.890) |
| Est 15 Years | | | $0.231^{***}$ |
| | | | (5.470) |
| Observations | 1,152 | 1,152 | 1,152 |
| Adjusted $R^2$ | 0.127 | 0.203 | 0.385 |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | |

Table 3: This table reports the standard and non-standard errors of portfolio returns keeping various specification choices fixed. Ratio refers to the ratio of the non-standard errors to the standard errors

| **Decision Type** | **Decision Value** | **SE** | **NSE** | **Ratio** |
|---|---|---|---|---|
| Weighting | Equal | 0.20 | 0.94 | 4.80 |
| | Value | 0.22 | 0.43 | 1.98 |
| Size Filter | No | 0.21 | 0.97 | 4.56 |
| | Yes | 0.20 | 0.58 | 2.93 |
| Financials Filter | No | 0.20 | 0.72 | 3.67 |
| | Yes | 0.21 | 0.77 | 3.61 |
| Utilities Filter | No | 0.20 | 0.71 | 3.45 |
| | Yes | 0.21 | 0.76 | 3.68 |
| Price Filter | No | 0.22 | 0.91 | 4.15 |
| | Yes | 0.19 | 0.62 | 3.21 |
| Age Filter | No | 0.22 | 0.78 | 3.57 |
| | Yes | 0.19 | 0.69 | 3.62 |
| Quantiles | 5 | 0.18 | 0.60 | 3.37 |
| | 10 | 0.23 | 0.81 | 3.45 |
| Training window size | 5 | 0.21 | 0.66 | 3.12 |
| | 10 | 0.20 | 0.76 | 3.71 |
| | 15 | 0.20 | 0.72 | 3.58 |

These results make it clear that certain methodological choices have a significant impact on the average performance of portfolios. Our analysis, however, can be extended beyond the averages. We now estimate the non-standard errors of the various portfolios observed and check how these methodological variations can be reduced. We calculate the standard and non-standard errors as defined in Section 2. The results of this analysis are contained in Table 3.

From the results in Table 3, we can conclude that the variation in returns due to methodological choices (i.e. non-standard errors) far exceeds the sampling variation (standard errors). Non-standard errors are frequently more than three times of the standard errors. Therefore, this source of variation should not be overlooked and if a study reports their findings using a particular set of choices, the overall results could vary to a large extent with only some seemingly innocent methodological choices. The quantitative results that we obtain are fairly intuitive. There is an extremely large variation in return for equally weighted portfolios, which can be curtailed by

Table 4: This table reports the standard and non-standard errors of value weighted portfolio returns after removing micro-cap stocks and considering only the XGBoost estimator. Ratio refers to the ratio of the non-standard errors to the standard errors

| Decision Type | Decision Value | Mean | SE | NSE | Ratio |
|---|---|---|---|---|---|
| Financials Filter | No | 0.90 | 0.20 | 0.31 | 1.53 |
| | Yes | 0.98 | 0.22 | 0.35 | 1.62 |
| Utilities Filter | No | 0.87 | 0.21 | 0.28 | 1.37 |
| | Yes | 1.01 | 0.21 | 0.40 | 1.91 |
| Price Filter | No | 0.92 | 0.21 | 0.37 | 1.77 |
| | Yes | 0.96 | 0.21 | 0.33 | 1.61 |
| Age Filter | No | 0.96 | 0.22 | 0.39 | 1.74 |
| | Yes | 0.92 | 0.20 | 0.32 | 1.64 |
| Quantiles | 5 | 0.76 | 0.18 | 0.20 | 1.12 |
| | 10 | 1.12 | 0.24 | 0.27 | 1.11 |
| Training window size | 5 | 1.02 | 0.23 | 0.31 | 1.38 |
| | 10 | 0.89 | 0.21 | 0.32 | 1.54 |
| | 15 | 0.92 | 0.19 | 0.28 | 1.45 |

giving more weight to larger stocks. Similarly, some variation across specifications can be reduced by preferring to filter out micro-caps and penny stocks. These results are not just useful for reducing method-induced variation, but also for enhancing the practical and real-time applicability of ML-based portfolio strategies (Avramov et al., 2023).

Therefore, the results highlight that there is a large variation in average returns for portfolios created from various paths created from methodological choices. At the same time, we also observe that certain choice paths may be disproportionately contributing to this phenomenon. We clearly observe that the choice of weighting, size and price filters, and the choice of ML method have a large impact on mean returns as well as the non-standard errors. Therefore, in the next set of results, we control for such obvious choices. In the results shown in Table 4, we eliminate micro-caps, consider value-weighted portfolios, and only the XGBoost estimator. Therefore, we control for the methodological return variation arising due to keeping smaller stocks in the sample and attaching higher weights to them. We also eliminate the variation due to the choice of ML estimator. We do all this to give a more realistic perspective on the size of the non-standard errors.

As per the results in Table 4, we observe a marked decline in the non-standard error

of our portfolios. However, despite the elimination of certain large-impact choices, a large amount of methodological variation still persists. Non-standard errors are still larger than the average standard errors. Some further filtering might reduce the ratio further, but the fact still looms large - non-standard errors are large enough to warrant our attention. Some choices could reduce their impact, but they are as important as the other source of variation, i.e. standard errors.

## 4 Conclusion

Studies using machine learning techniques for return forecasting have shown a lot of promise. However, as in empirical asset-pricing, researchers have to make many choices revolving around the sampling and estimation of forecasting models.

We quantify the impact of methodological choices on the performance of machine-learning-based strategies. Results from more than 1152 choice combinations show that there is a sizeable variation in the average returns of ML strategies. The usage of value-weighted portfolios with size filters can curb a good portion of this variation, but cannot eliminate it. So, what is the solution to non-standard errors? Studies in empirical asset pricing have proposed various solutions. While Soebhag et al. (2023) suggests that researchers can show a distribution of outcomes across major specification choices, Walter et al. (2023) argues in favour of reporting the entire distribution across all specifications.

While we agree with reporting results across variations, we would also advise against a one-size-fits-all solution for this issue. Despite an extensive computation burden, It is possible to compute and report the entire distribution of returns for characteristic-sorted portfolios as in Walter et al. (2023). However, when machine-learning methods are used, reporting the full distribution is likely to impose an extreme computational burden on the researcher. Although a full distribution is more informative than a partial one, the costs and benefits of both choices need to be evaluated before giving generalized recommendations. In future drafts of this paper, we intend to explore additional ways to control for methodological variation while imposing a modest burden on the researcher. At present, our recommendations tilt in favor of first identifying selected high-impact choices (e.g. weighting and size filters) on a smaller-scale analysis. Researchers can then, at the very least, report variations of results across such high-priority specifications, while keeping the rest optional.

## References


Avramov, D., S. Cheng, and L. Metzker (2023, May). Machine Learning vs. Economic Restrictions: Evidence from Stock Return Predictability. *Management Science 69*(5), 2587–2619.

Azevedo, V., G. S. Kaiser, and S. Mueller (2023, September). Stock market anomalies and machine learning across the globe. *Journal of Asset Management 24*(5), 419–441.

Bali, T. G., H. Beckmeyer, M. Moerke, and F. Weigert (2023). Option return predictability with machine learning and big data. *The Review of Financial Studies 36*(9), 3548–3602. Publisher: Oxford University Press.

Banz, R. W. and W. J. Breen (1986, September). Sample-Dependent Results Using Accounting and Market Data: Some Evidence. *The Journal of Finance 41*(4), 779–793.

Bianchi, D., M. Büchner, and A. Tamoni (2021). Bond risk premiums with machine learning. *The Review of Financial Studies 34*(2), 1046–1089. Publisher: Oxford University Press.

Blitz, D., M. X. Hanauer, T. Hoogteijling, and C. Howard (2023). The term structure of machine learning alpha. *Available at SSRN*.

Breiman, L. (2001, October). Random Forests. *Machine Learning 45*(1), 5–32.

Cakici, N., C. Fieberg, D. Metko, and A. Zaremba (2023a). Machine learning goes global: Cross-sectional return predictability in international stock markets. *Journal of Economic Dynamics and Control 155*, 104725. Publisher: Elsevier.

Cakici, N., C. Fieberg, D. Metko, and A. Zaremba (2023b). Machine learning goes global: Cross-sectional return predictability in international stock markets. *Journal of Economic Dynamics and Control 155*, 104725. Publisher: Elsevier.

Cakici, N., S. J. H. Shahzad, B. Będowska-Sójka, and A. Zaremba (2024). Machine learning and the cross-section of cryptocurrency returns. *International Review of Financial Analysis 94*, 103244. Publisher: Elsevier.

Chen, A. Y. and T. Zimmermann (2021). Open source cross-sectional asset pricing. *Critical Finance Review, Forthcoming*.

Chen, T. and C. Guestrin (2016, August). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA, pp. 785–794. ACM.

Coqueret, G. (2023). Forking paths in empirical studies. *Available at SSRN 3999379*.

Fabozzi, F. J. and M. L. d. Prado (2018, November). Being Honest in Backtest Reporting: A Template for Disclosing Multiple Tests. *The Journal of Portfolio Management 45*(1), 141–147. Company: Institutional Investor Journals Distributor: Institutional Investor Journals Institution: Institutional Investor Journals Label: Institutional Investor Journals Publisher: Portfolio Management Research Section: Quantitative Finance.

Fieberg, C., S. Günther, T. Poddig, and A. Zaremba (2024). Non-standard errors in the cryptocurrency world. *International Review of Financial Analysis 92*, 103106. Publisher: Elsevier.

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

Harvey, C. R. (2017, August). Presidential Address: The Scientific Outlook in Financial Economics. *The Journal of Finance 72*(4), 1399–1440.

Harvey, C. R. and Y. Liu (2020, October). False (and Missed) Discoveries in Financial Economics. *The Journal of Finance 75*(5), 2503–2553.

Leamer, E. and H. Leonard (1983). Reporting the fragility of regression estimates. *The Review of Economics and Statistics*, 306–317. Publisher: JSTOR.

Leamer, E. E. (1983). Model choice and specification analysis. *Handbook of econometrics 1*, 285–330. Publisher: Elsevier.

Menkveld, A. J., A. Dreber, F. Holzmeister, J. Huber, M. Johannesson, M. Kirchler, S. NEUSÜß, M. Razen, U. Weitzel, D. Abad-Díaz, M. M. Abudy, T. Adrian, Y. Ait-Sahalia, O. Akmansoy, J. T. Alcock, V. Alexeev, A. Aloosh, L. Amato, D. Amaya, J. J. Angel, A. T. Avetikian, A. Bach, E. Baidoo, G. Bakalli, L. Bao, A. Barbon, O. Bashchenko, P. C. Bindra, G. H. Bjønnes, J. R. Black, B. S. Black, D. Bogoev, S. B. Correa, O. Bondarenko, C. S. Bos, C. Bosch-Rosa, E. Bouri, C. Brownlees, A. Calamia, V. N. Cao, G. Capelle-Blancard, L. M. C. Romero, M. Caporin, A. Carrion, T. Caskurlu, B. Chakrabarty, J. Chen, M. Chernov, W. Cheung, L. B. Chincarini, T. Chordia, S. Chow, B. Clapham, J. Colliard, C. Comerton-Forde, E. Curran, T. Dao, W. Dare, R. J. Davies, R. D. Blasis, G. F. D. Nard, F. Declerck, O. Deev, H. Degryse, S. Y. Deku, C. Desagre, M. A. V. Dijk, C. Dim, T. Dimpfl, Y. J. Dong, P. A. Drummond, T. Dudda, T. Duevski, A. Dumitrescu, T. Dyakov, A. H. Dyhrberg, M. Dzieliński, A. Eksi, I. E. Kalak, S. T. Ellen, N. Eugster, M. D. D. Evans, M. Farrell, E. Felez-Vinas, G. Ferrara, E. M. Ferrouhi, A. Flori, J. T. Fluharty-Jaidee, S. D. V. Foley, K. Y. L. Fong, T. Foucault, T. Franus, F. Franzoni, B. Frijns, M. Frömmel, S. M. Fu, S. C. Füllbrunn, B. Gan, G. Gao, T. P. Gehrig, R. Gemayel, D. Gerritsen, J. Gil-Bazo, D. Gilder, L. R. Glosten, T. Gomez, A. Gorbenko, J. Grammig, V. Grégoire, U. Güçbilmez, B. Hagströmer, J. Hambuckers, E. Hapnes, J. H. Harris, L. Harris, S. Hartmann,

J. Hasse, N. Hautsch, X. T. He, D. Heath, S. Hediger, T. Hendershott, A. M. Hibbert, E. Hjalmarsson, S. A. Hoelscher, P. Hoffmann, C. W. Holden, A. R. Horenstein, W. Huang, D. Huang, C. Hurlin, K. Ilczuk, A. Ivashchenko, S. R. Iyer, H. Jahanshahloo, N. Jalkh, C. M. Jones, S. Jurkatis, P. Jylhä, A. T. Kaeck, G. Kaiser, A. Karam, E. Karmaziene, B. Kassner, M. Kaustia, E. Kazak, F. Kearney, V. V. Kervel, S. A. Khan, M. K. Khomyn, T. Klein, O. Klein, A. Klos, M. Koetter, A. Kolokolov, R. A. Korajczyk, R. Kozhan, J. P. Krahnen, P. Kuhle, A. Kwan, Q. Lajaunie, F. Y. E. C. Lam, M. Lambert, H. Langlois, J. Lausen, T. Lauter, M. Leippold, V. Levin, Y. Li, H. Li, C. Y. Liew, T. Lindner, O. Linton, J. Liu, A. Liu, G. Llorente, M. Lof, A. Lohr, F. Longstaff, A. Lopez-Lira, S. Mankad, N. Mano, A. Marchal, C. Martineau, F. Mazzola, D. Meloso, M. G. Mi, R. Mihet, V. Mohan, S. Moinas, D. Moore, L. Mu, D. Muravyev, D. Murphy, G. Neszveda, C. Neumeier, U. Nielsson, M. Nimalendran, S. Nolte, L. L. Norden, P. O'Neill, K. Obaid, B. A. Ødegaard, P. Östberg, E. Pagnotta, M. Painter, S. Palan, I. J. Palit, A. Park, R. Pascual, P. Pasquariello, L. Pastor, V. Patel, A. J. Patton, N. D. Pearson, L. Pelizzon, M. Pelli, M. Pelster, C. Pérignon, C. Pfiffer, R. Philip, T. Plíhal, P. Prakash, O. Press, T. Prodromou, M. Prokopczuk, T. Putnins, Y. Qian, G. Raizada, D. Rakowski, A. Ranaldo, L. Regis, S. Reitz, T. Renault, R. W. Renjie, R. Reno, S. J. Riddiough, K. Rinne, P. Rintamäki, R. Riordan, T. Rittmannsberger, I. R. Longarela, D. Roesch, L. Rognone, B. Roseman, I. Roşu, S. Roy, N. Rudolf, S. R. Rush, K. Rzayev, A. A. Rzeźnik, A. Sanford, H. Sankaran, A. Sarkar, L. Sarno, O. Scaillet, S. Scharnowski, K. R. Schenk-Hoppé, A. Schertler, M. Schneider, F. Schroeder, N. Schürhoff, P. Schuster, M. A. Schwarz, M. S. Seasholes, N. J. Seeger, O. Shachar, A. Shkilko, J. Shui, M. Sikic, G. Simion, L. A. Smales, P. Söderlind, E. Sojli, K. Sokolov, J. Sönksen, L. Spokeviciute, D. Stefanova, M. G. Subrahmanyam, B. Szaszi, O. Talavera, Y. Tang, N. Taylor, W. W. Tham, E. Theissen, J. Thimme, I. Tonks, H. Tran, L. Trapin, A. B. Trolle, M. A. Vaduva, G. Valente, R. A. V. Ness, A. Vasquez, T. Verousis, P. Verwijmeren, A. Vilhelmsson, G. Vilkov, V. Vladimirov, S. Vogel, S. Voigt, W. Wagner, T. Walther, P. Weiss, M. V. D. Wel, I. M. Werner, P. J. Westerholm, C. Westheide, H. C. Wika, E. Wipplinger, M. Wolf, C. C. P. Wolff, L. Wolk, W. Wong, J. Wrampelmeyer, Z. Wu, S. Xia, D. Xiu, K. Xu, C. Xu, P. K. Yadav, J. Yagüe, C. Yan, A. Yang, W. Yoo, W. Yu, Y. Yu, S. Yu, B. Z. Yueshen, D. Yuferova, M. Zamojski, A. Zareei, S. M. Zeisberger, L. Zhang, S. S. Zhang, X. Zhang, L. Zhao, Z. Zhong, Z. I. Zhou, C. Zhou, X. S. Zhu, M. Zoican, and R. Zwinkels (2024, June). Nonstandard Errors. *The Journal of Finance 79*(3), 2339–2390.

Newey, W. K. and K. D. West (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica 55*(3), 703. Publisher: JSTOR.

Rasekhschaffe, K. C. and R. C. Jones (2019, July). Machine Learning for Stock Selection. *Financial Analysts Journal 75*(3), 70–88.

Sala-I-Martin, X. X. (1997). I Just Ran Two Million Regressions. *The American Economic Review 87*(2), 178–183. Publisher: American Economic Association.

Shumway, T. (1997, March). The Delisting Bias in CRSP Data. *The Journal of Finance 52*(1), 327–340.

Soebhag, A., B. Van Vliet, and P. Verwijmeren (2023). Non-standard errors in asset pricing: Mind your sorts. *Available at SSRN 4136672*.

Walter, D., R. Weber, and P. Weiss (2023). Non-standard errors in portfolio sorts. *Available at SSRN 4164117*.