AIPA's International Journal on Artificial Intelligence: Bridging Technology, Society and Policy
ISSN: 3062-097X
Published: 07 December 2024

Reliability And Outcome Bias Issues In AI-Driven Forecasting Practices

Mehmet Beceren[1*]

[1]**Queen's University Smith School of Business, Kingston, ON, K7L 3N6; FORA-Invest Research, Oakville, ON; ORCID: 0009-0000-0238-6304)**

ORIGINAL RESEARCH PAPER

Abstract

Amid all the hype around the economic potential of AI technologies, there is a growing risk of data analysis overkill in many applications. That risk is particularly high for the forecasting and decision-making models being proposed in social contexts such as economic policy, financial investment, and corporate decisions. Common research practices in those areas keep focusing on incidents of statistical discoveries. They omit the substantial reliability issues stemming from the nature of the data that offers very limited 'learning potential' for the machine learning (ML) algorithms. In this paper, I focus on the use of ML algorithms applied to such forecasting problems. I illustrate the reliability issues with a detailed example that builds a stock investment strategy by using the XGBoost algorithm on a large data set. The example demonstrates how easy it is to discover seemingly interesting random patterns when we fit over-parameterized models on historical data. The results also offer practical methods to investigate the statistical flukes and the reliability issues that are concealed by complex algorithms of artificial intelligence being blended with natural human ignorance, as seen in popular practice.

Keywords: forecasting, reliability, machine learning, asset pricing, factor investing

1 Introduction

> "*It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of light, it was the season of darkness, it was the spring of hope, it was the winter of despair*". Tale of Two Cities by Charles Dickens

This famous opening line of the Charles Dickens classic, *Tale of Two Cities*, works perfectly to encapsulate the main theme of this article in a nutshell. The simple and timeless language of the novel fits quite well to our data-obsessed times.

> "*It is the best of times, it is the worst of times, it is the age of artificial intelligence, it is the age of human ignorance, it is the epoch of data analytics, it is the epoch of statistical deception.*"

In the current digital age, there is a euphoric race both among businesses and academics to showcase the latest machine learning (ML) applications in their own practice areas. We see an exponential growth in ML-driven research output and commercial applications that utilise increasingly complex predictive models with ever-larger data sets. Amid all the buzz around the economic potential of artificial intelligence (AI) technologies, however, there is also a growing risk of data analysis overkill in many cases. The rush to catch up with the self-fulfilling 'AI revolution' wave is inevitably generating misused, misguided implementations alongside many fascinating products. That risk is particularly high for the forecasting and decision-making models being proposed in social contexts such as economic policy, financial investment, corporate strategy and such.

In this paper, I focus on the use of ML algorithms applied to forecasting problems. I discuss the unique nature and the limitations of historical data sets that have stochastic state-and-time dependent variables. I illustrate the specific issues with detailed examples from the financial investment strategy applications.

The main sources of concern about the excessive use of ML techniques to build decision models are as follows:

OPEN ACCESS
AIPAJ Vol:1, Issue:1

***Corresponding author**
mehmet.beceren@queensu.ca

***Submitted** 22 June 2024*

***Accepted** 15 November June 2024*

Citation
Beceren, M. (2024). Reliability And Outcome Bias Issues In AI-Driven Forecasting Practices, In AIPA's International Journal on Artificial Intelligence: Bridging Technology, Society and Policy, (Vol. 1, Number 1, pp. 25–40. DOI: 10.5281/zenodo.11200944

1. A unique sequence of historical events caused by incidental patterns of stochastic factors, and complex confounding effects, do not provide useful data sets that are sufficient to make reliable inferences about the future. In other words, unlike many successful applications such as image recognition, complex pattern discoveries within historical data sets, may not amount to 'learning' or 'intelligence' of any sort.

2. Although the Train-Test-Validation cycle of the ML algorithms may generate incidents of attractive back-test results (i.e. performance validations) on historical data, the relation between the performance metrics and future reliability may be highly uncertain.

3. In cases where (1) and (2) are true, there are significant reliability and outcome bias issues in ML-driven models. A forecasting solution that looks encouraging with historical data, may easily be an *over-fitted* fluke driven by a lucky draw from a large random set.

Surprisingly, neither academics nor professionals in social sciences tend to sufficiently address these serious issues. The hype to assign a flashy "AI" label on new products seems to trump the obvious reliability challenges. Probably fascinated by the speed and efficiency of ML algorithms, the data analysts seem to ignore the significant likelihood of making incidental, lucky discoveries with big data. Also, they tend to forget that a longer history, occurred and evolved with unique circumstances in time, does not necessarily mean a bigger data set with relevant and useful information.

The following sections will discuss the reliability risks in further detail along with some examples from the finance literature. At this point, however, it is probably a good idea to offer a bit more clarification about the concepts mentioned above for the non-expert reader.

1.1 Data mined flukes versus reliable insights

To understand the outcome bias and statistical flukes found in historical data analysis, let's consider an extreme case where the target variable (i.e predicted or forecast variable) is completely random. Assume that you are the manager of a company named Lucky Bets Co. You believe in luck and in lucky people. You are in the gambling business, but you do not place bets on games. Instead, you place bets on lucky people. You provide funding for the gamblers that you think are lucky to win at the roulette table in return for a large share of the prizes they win.

The skillful data analysts of Lucky Bets Co. collect a large historical data set on many attributes of the addicted roulette players. The data set includes the players' winning percentage over the past 5 years, amount of money they lost, age, height, profession, post code, shoe size, hair color, first letter of their names, star sign, and many others. The analysts divide the data set into Train, Test and Validation samples, and then let the ML algorithms run over-parameterized deep learning models, as they always do. After millions of iterations, the analysts provide a combination of attributes that predict a higher probability of winning at the roulette table. The results are confirmed in the Validation (hold-out) sample as well. All standard statistical measures check within the Test and Validation sub-samples.

What would you do? Would your expectation of winning probability change for the people with the right attributes? Assuming everyone plays the same game with the same odds, would you bet on the people with "statistically proven" success? Are there lucky characteristics, or lucky data analysts here?

Your betting decision actually does not matter. It will not change the odds of winning one way or another. The data analysts did not do much more than wasting electricity. They were lucky. Also, it was almost inevitable that they would find a fluke that works after so many iterations over countless combinations of gambler attributes. The historical results, no matter how statistically significant they may look, provide no guidance for the future outcomes that are completely random. The analysts just documented an observation bias - a lucky historical outcome with no implication for the future. That is because each roulette run is an independent random event by construct.

On the other hand, it may actually be a good strategy to go along with the model and promote it as the new, cutting-edge AI-Powered innovation by Lucky Bets Co. If, somehow, it catches another lucky episode, it may bring extra fame and fortune. (Actually, there are online betting companies, especially in sports betting, that offer AI models for their customers. See examples such as *DeepBetting*, *BetIdeas* or *Infinity Sports AI* among others.)

Typically, when there is a proposed forecast model, or a decision method, we are likely to see some instances of out-of-sample performance metrics as the key results. An instance of out-of-sample test is considered sufficient to prove the suggested model's worth. The reliability risk and potential 'observation bias' originating

from the iterative data mining embedded in over-parameterized ML models are mostly downplayed. As a result, the real important question is mostly left unanswered.

Given that we are able to find some model that performed well in the past, how confident are we that the model will provide significant performance in the future as well? What is the correlation of the actual implemented results with the (out-of-sample) past performance that we could dig out by sifting through the data?

In the case of Lucky Bets Co., we know the answer. The correlation is zero. If we keep repeating the exercise of finding new hidden patterns with strong past performance, by utilising more and more data, and then we implement each model as a separate AI-powered betting strategy, we surely will find out that the documented past results have no relevance for future outcomes. Such an analysis would serve as the proper back-testing experiment to provide some guidance about the reliability of the methods.

Those experiments almost never show up in the results of ML-driven forecasting research, especially in the social science fields such as economics and finance. Both academic researchers and professionals keep showing instances of statistical discoveries, instead. Their common audience usually cannot distinguish the lucky coincidences hidden behind the complex and automated algorithms.

The computational power of the ML algorithms help the empirical researchers with the fast discovery of interesting patterns, but the findings might be just an '*observation bias*' - a fluke of the unique set of circumstances that might not repeat ever again. Therefore, when we try to import the predictive AI technologies to forecasting practices, one of the first questions to ask has to be: "*How similar is my case to Luck Bets Co.?*"

Many examples of empirical research output that are being promoted with sparkling AI labels might not be far from just another Lucky Bets exercise. It is common to find similar examples, especially in fields that rely on non-repeatable, state-and-time dependent data. Just to mention a few, Berman et al. (2021) [1], presents a model that integrates big data analytics with strategic planning to optimize business decisions; Lee and Chen (2020) [2] presents a machine learning model that predicts both employee success and retention; Chen and Guestrin (2016) [3] predicts political instability with ML models fitted onto social media data, and many others. In each study, we see some contemporaneous covariance among variables being documented with no in-depth discussion about cross-validation and reliability issues originating from particular methods and data samples used.

Another example, Erel et al. (2021) [4], presents results of decision tree models to select directors for corporate executive boards. The target variable used is "director success" which is some complex proxy measure constructed with authors' subjective discretion. It includes ad hoc indicators of shareholder popularity and company profitability. The ML algorithms run an over-parameterized decision tree model on a predetermined training sample and a fixed test sample. The model iterates over tens of different personal attributes, from gender and age, to the name of the university that the director graduated from. There is no cross-validation across different periods, industries, etc. There is no proper validation experiment over time either. The incident of the statistical results are particular to a very narrowly defined data construction process.

To find some interesting-looking pattern in large data sets does not require much skill since we have the technology to automatically iterate over pretty much countless parameter combinations. Those empirical research articles, and many other similar work, are arguably not that far away from the Lucky Bets case. Although the publications succeed in uncovering intriguing incidents of empirical results, future reliability of the findings, as a useful forecasting model, is a wide open question.

Historical data sets used for forecasting models in social contexts usually do not offer the breadth for proper cross-validation tests. After all, we have only one trail of the actual history. Therefore, AI methods that are employed successfully in other areas, may be unsuitable, or misleading, due to the irreducible over-fitting risk originating from the nature of the data sets. Quick and lazy ML applications with historical data require scrutiny within their own context since the standard data validation methods are mainly not feasible.

1.2 A special case: Financial asset pricing and investment strategy applications

Finance has been at the forefront of digital automation and the commercial use of AI technologies. Financial industry operates on an extremely digitized platform that produces immense amount of data, and the data universe is mostly accessible for analysis. Data collection is relatively easy and straightforward. Financial industry employees, especially on the trading and investment side, tend to be highly skilled in data analysis and coding practices. At the same time, the potential reward of successful forecasting models can be very high and fast especially in the trading and investment world.

In addition to the general economic backdrop that motivates the use of ML models in finance, the academic literature also provides some extra justification for the use of sophisticated predictive models in this field. For example, the investment management industry makes use of models inspired by the academic asset pricing literature. Contemporary empirical research in this field has developed around the Arbitrage Pricing Theory (APT), introduced by Ross (1976) [5], and the Stochastic Discount Factor concept, introduced by Merton (1973) [6], that lay out the framework for the empirical inquiries into the driving factors of financial asset returns. The seminal work by Fama and French (1992, 1993) [7],[8] and a large body of empirical work that followed the same path into the inquiry of asset returns, built a cultural tradition that is baked into the contemporary curriculum of finance education. The highly-regarded Chartered Financial Analyst (CFA) program also teaches the APT and related concepts that underpin the empirical inquiries into historical data to search for the drivers (factors) of asset returns.

The basic idea is that the financial asset returns are determined by their sensitivity to (potentially many) risk factors that the agents trade in the market place. It sounds like an axiomatic statement that opens up a wide gate for the inquiry of those elusive factors.

The complex and efficient predictive machinery offered by the recent developments in AI technology are welcomed as a powerful tool to work on the eternal questions of the investment industry and the asset pricing academics: *What drives the differences in asset returns? What should be the decision criteria to choose the assets to invest for the short or the long term?*

To answer those questions, quantitative finance professionals and academics dedicate a great portion of their work to building predictive models for the asset return dynamics. Common empirical research practice starts with an investigation of the so-called factors that show some covariance with the cross-sectional variance of asset returns in hand. Once the candidates for useful factors and trading signals are found, they are put into a back-testing process to validate their historical success. The instances of out-of-sample back-test results achieved over a selected period is usually considered as a sufficient experiment result. Reliability is mostly left out of the discussion.

With the advances in data access and computer power, the statistical discoveries became rather easy and fast. Sequentially, the number of academic publications showcasing the discovery of new factors started to grow rapidly during the early years of this century. From economic and financial indicators, to eccentric sentiment and risk measures, numerous variables are thrown into predictive models with the hope of finding some covariance patterns. The finance professionals started to implement such models for portfolio construction and proprietary trading practices at an accelerating pace, as well. By the time we reached 2010s, the asset pricing literature became a 'factor zoo' as famously coined by Cochrane (2011) [9]. The criticism and warnings about the scientific quality of the empirical findings began to accumulate.

The critics highlighted two key observations. One, the published articles were presenting obviously over-fitted models that did not pass the statistical hurdle tests and the test of time. Two, the investment strategies based on the suggested factors mostly failed to deliver returns documented in their back-tests. In other words, the real out-of-sample tests proved that neither the predictive models nor the underlying theory was able to deliver a decent reliability over time.

The published statistical results were not necessarily wrong or careless, however. The issue was that the suggested models were not far from our Lucky Bets Co. example, again. People put too much faith in the instances of pattern discoveries driven from over-simplified models. Even the factors suggested by Nobel Prize winning Eugene Fama and Ken French's work failed to repeat the documented patterns consistently, once they were implemented as real investment strategy products. See Carhart (1997), Fama, French and Carhart (2000), Fama and French (2015) [10, 11, 12] for more detail on that point.

As a result of humbling real-world validation experiences in the financial markets, the discussions on the potential uses of ML-driven or other type of predictive models started to shift from euphoria to skepticism, especially over the past 10 years. At this point, we can probably say that finance is more advanced in the discussions about reliability compared to other social science fields.

The discussions are evolving in three main paths. The first path can be called the 'scientific quality' argument. Studies such as Bailey and Prado (2013, 2014), Prado (2020), Harvey et al. (2016), [13, 14, 15], [16, 17] present strong arguments about the 'data mining' and 'over-fitting' issues. They discuss the rampant use of statistical overkill and careless back-test practices spoiled by the ease of access to computational tools and large data sets. The criticism raised by Prado and Harvey is mainly about the errors, tricks and and biases in statistical inference. They are valid and crucial points that highlight the risk of false discoveries and wrong inferences made in common research practices.

However, the failures of the forecasting models in this field are not necessarily driven by the lack of diligence in statistical analysis. It is driven by the fact that there is an irreducible reliability issue caused by the natural instability of the system dynamics. To argue for scientific quality of the predictive models applied to naturally unpredictable dynamics is somewhat redundant. After all, it is impossible to determine the causes of model failure with confidence when the model is too simple relative to the stochastic complexity of the system in hand. As the past research experience showed time and time again, no matter how robust your statistical results may be, the estimated model may fail to perform, or become irrelevant, simply because of the evolving complexity of the system not being captured by the available historical data.

Second path can probably be categorized as the 'benign over-fitting' effort. Studies such as Kelly et al. (2022) [18, 19] do not find the risk of over-fitting as an impediment to ML-based iterative search for hidden patterns. Instead, they try to develop the machinery that let over-parameterized, over-fitted models to automatically iterate towards a rather distilled form. They also let the ML algorithms to adjust over time, and over different states, and also let the algorithms discover those adjustment rules independently from the data. This line of research focuses on methods to distill signals without being limited by theory, or any other priors. It is probably a step in the right direction with a powerful inspiration, but reliability is still mostly missing in the discussion. Instances of good-looking back-tests are presented without a demonstration of how reliably the complex models might perform relative to simpler decision rules over time.

The third path suggests an alternative use case for the ML algorithms. The work by Chean and Zimmerman (2020) , and Chen and Valikov (2021) [20, 21] embraces ML-powered intentional data-mining to investigate the reliability of the models proposed by the 'factor zoo' literature mentioned above. The approach is a leap from simply documenting another discovery of factors towards an analysis of real out-of-sample performance. With a multitude of different data-mined correlations that can easily present some historical performance, this line of work aims to establish a benchmark for the value and usefulness of the models that claim to have some prediction power.

I think Chen's work is an example of how the AI technologies can bring a significant disruption to social sciences and forecasting practices. By allowing the fast and automated search algorithms, ML models can help us to devise tools to help distinguish a humble analysis that provides insights to highly complex and fluid stochastic systems from a statistical fluke published with a dose of confirmation bias and academic hubris.

Meanwhile, although similar discussions happen in parts of the investment industry, the commercial pressure to roll out generic commercial products with a flashy AI-name continues. Take the "AI-labeled" exchange traded fund (ETF), QRFT - QRAFT AI Enhanced US Large Cap ETF, for example. This ETF relies on "AI-powered models" which are based on some back-tested historical correlations - not some "intelligence" gained by learning from very large big data sets as we see in other fields. As seen in the Figure 1, there is no convincing performance of any sort. The performance over the benchmark index converges to zero as you would expect from any Lucky Bets exercise. In academia, as well, we can observe an intellectual inertia to keep producing those incidental back-test results. The publication rate of such research will inevitably fade away as their value-added is tested over time.

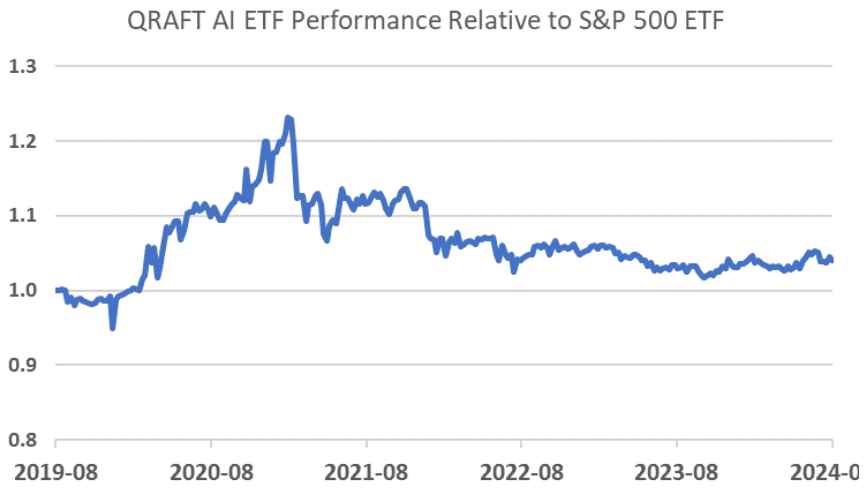

Figure 1. AI-Powered Large Cap US ETF Performance; *Source: www.qraftaietf.com*

Motivated by the contributions of all three paths followed by recent finance literature discussed above, the following section presents an example of an investigation into an ML-driven method applied to asset returns.

First, by using historical data on stocks and company characteristics, I run a decision-tree model (XGBoost) to intentionally data-mine the factors that distinguish the Winner (high future return) and Loser (low future return) stocks - similar to our Lucky Bets case. I demonstrate how easy it is to document some seemingly successful back-test when you are not much concerned about cross-validation. Then, I run a series of investigations to discuss how similar the case could be to the Lucky Bets scenario.

I do not use the ML techniques to show how we can predict Winners and Losers in the stock market. Instead, I utilise the power of ML to show how reliable the employed data and methods might be for the specific case in hand.

2 Material and Methods

Let's assume we have a problem of building an ML-based stock selection method that can possibly be turned into an ETF product similar to the one mentioned above. However, the financial literature does not offer much help about the predictors of stock returns. Although there are some obvious common sense approaches to portfolio construction and investment, there is no formula to predict which stocks will outperform the others over a certain period, say, the next 3 months or 2 years. Actually, there cannot be a formula because, if there was one, it would be instantly exploited and vanish, anyway.

The markets facilitate exchange of expected risks and returns that fluctuate according to perceived opportunities and costs that vary across numerous agents over time and economic conditions. Incidental clusters of those expectations cause demand-supply imbalances to move the asset prices. Additionally, when the underlying assets deliver unexpected positive or negative economic performance, share prices adjust so as to remain consistent with changing conditions.

Although financial theory does not offer a magic formula, at least it provides the framework that allows empirical investigations for the elusive, incidental or persistent risk factors that drive returns.

The equation for the expected asset return $R_{i,t}$ for the asset i at time t is given by:

$$E[R_{i,t+1}] = \Gamma_t(\beta_{i,t} \cdot X_t)) \tag{1}$$

where:

- $R_{i,t+1}$ is the return to be realized at time $t+1$, E is the expectation operator,
- $\beta_{i,t}$ is the $N \times 1$ vector of exposure of asset i to the observed N factors X_t,
- X_t is the $(1 \times N)$ vector of factors that are assumed to affect expectations,
- Γ_t is the time-specific function that translates observed factors to retun expectations

Here, one can think of X_t as the set of themes and criteria that influences asset return expectations and portfolio preferences at a point in time. For example, they may be a popular theme such as AI to drive growth expectations for the share price of Nvidia lately. The exposure $\beta_{i,t}$ of Nvidia to the AI theme may be high while for a company such as Alcoa which is in the business of metal mining globally, $\beta_{i,t}$ may be zero. One can think of $\beta_{i,t}$ as traffic lights switching on and off over time differently for each stock as themes, risks and investors' preferences evolve.

The issue is that we do not know any of those parameters in that simple abstraction (1). We have some idea about what the investors generally consider, maybe factors such as profitability, volatility etc., but we have no idea how those considerations might translate into return performance at a point in time. We would like to believe that we have some intuitive list of what X_t could consist of, but we do not have a clean method of measurement either. Therefore, equation (1) does not tell us anything other than 'whatever works!' offers no insights. (That pretty much sums up the field of asset pricing in finance.)

All we have is the historical realizations of $R_{i,t}$ and a data set of factors X_t that we imagine, and hope, will show some covariance with future returns to help us distinguish the Winners and Losers. So, as one can easily see, the problem in hand is not much different from the Lucky Bets scenario discussed earlier.

Our case is probably a very good example of potential use cases of AI to solve complex problems without a specific formula. We observe some phenomena that is driven by complex interactions of unknown set of factors. We hope that the computational technology will be able to sift through huge data sets to generate useful predictions although we are not able to identify what exactly drives those predictions. Image

recognition with deep neural networks is such a process. We cannot tell how exactly the image recognition works, but we see that computer algorithms trained on big-enough data sets can accumulate the cognitive experience to generate impressively accurate predictions. A deep learning model trained on millions of X-ray images, for example, comes close to obtaining a life-time experience of a doctor. That is made possible by being exposed to a very large number of instances of a well-defined problem.

To have access to *'the instances of a well-defined problem'* is the key issue that distinguishes forecasting from other problems. As we increase the size of our data set, by extending the history for example, we do not necessarily accumulate the instances to learn from. The phenomena that we register in our data sets are mostly the outcome of instances of unique or temporarily relevant circumstances. That is why, with historical finance data, we do not see the 'double descent' phenomenon that is remarkably demonstrated by Belkin (2021) [22]

Alonso and Sonam (2023) [23] applies Belkin's (2021) [22] methods to financial return data set and shows that the learning accuracy rate does not improve with larger data sets with more parameters. Alonso and Sonam (2023) [23] formally experiments with the financial data sets and documents how the historical data sets fail to show any potential for 'double descent'.

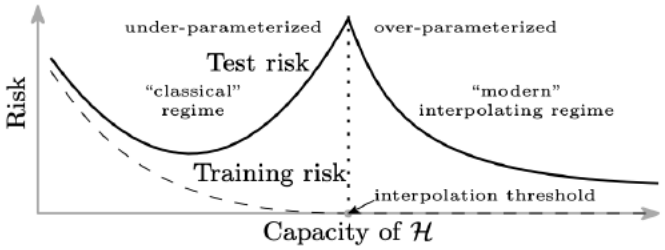

Figure 2. Double-descent of over-parameterized ML models shown in Belkin (2021)

In our case, we have a similar data set with (very) limited learning capacity. In our modeling exercise, we need to humbly accept that fact, and try to analyze what we can distill from the data set. As discussed in the earlier section, the main argument and motivation of this paper is the lack of such approaches in common ML-driven forecasting practices. There is too much focus on the instances of statistical findings, and too few discussion about how much luck is involved in those findings.

Our data set is the same as the one used in Guida (2020) [24]. The data is available through the book's Github. We have monthly data on the Total Return of 1212 global stocks over a 20-year period from 1998 to 2019. All the stock characteristics (features) to be used as predictors are scaled and normalised and they are ready to be used in ML algorithms. Not all stocks are alive throughout the 20 years. Some vanish, others emerge, as they always do, in the data. Therefore, we have an unbalanced panel of cross-sectional, time-series data with over 208K rows (roughly [20 year x 12 months x 1000 stocks]).

Along with the stock returns, there are also 93 different company characteristics such as valuation ratios, past returns, past volatility, accounting measures of profitability, growth, debt, capital expenditures, and many other similar variable with seemingly relevant economic measures. Of course, we do not know whether any of these variables make any reliable predictor of Winner or Loser stocks at any time. Although the variables seem to have financially meaningful labels, they are not necessarily different from any random number in relation to their predictive value for future stock returns.

Our prior is that we have some function given in (1) that will partly reveal itself in the large data set in hand. The data is aligned such that a model can be fitted as:

$$R_{i,t+1} = \Gamma_t'(\beta_{i,t}' \cdot Z_t) \ + \epsilon_{i,t+1} \tag{2}$$

where:

- $R_{i,t+1}$ is the return to be realized at time $t+1$,
- $\beta_{i,t}'$ is the estimate of exposures of asset i to the observed N factors Z_t,
- Z_t is the $(1 \times N)$ vector of factors that we have in hand with no causal relation with the returns, necessarily

- Γ'_t is the time-specific estimated function that translates observed factors to future returns observed

We transform the problem to the following form:

$$rank[R_{i,t+1}] = rank[\Gamma'_t(\beta'_{i,t} \cdot Z_t)] + \phi_{i,t+1} \tag{3}$$

because we are interested in the Rank of the future returns across stocks at a point in time. We set $R_{i,t+1}$ as the Next 3-Month Return. For example, in 2009-December, we would like to predict the Rank of returns over the 3 months from 2010-Jan to 2010-Mar. At each point in time (i.e. each Month in the data set), we define the top 80% as Winners = 1, and the bottom 20% as Losers (Winners = 0).

To fit a tree-based model, we can use the XGBoost (Extreme Gradient Boosting) algorithm. XGBoost builds an ensemble of trees sequentially, where each tree corrects the errors of the previous ones by focusing on the hardest-to-predict cases. The algorithm incorporates regularization to prevent over-fitting. It is popular in categorization (1 vs. 0) problems. The model output includes decision trees similar to the Figure 3 below.

As a start, let's pick a small portion of the large data set. Let's take the first 3 years as the *Train*, and pick the 3 months immediately after the *Train*, as the *Test* sample. Our hope is that the model will train on the past 36 months as the 'most relevant' period to forecast the Winner and Loser stocks in the next 3-month period.

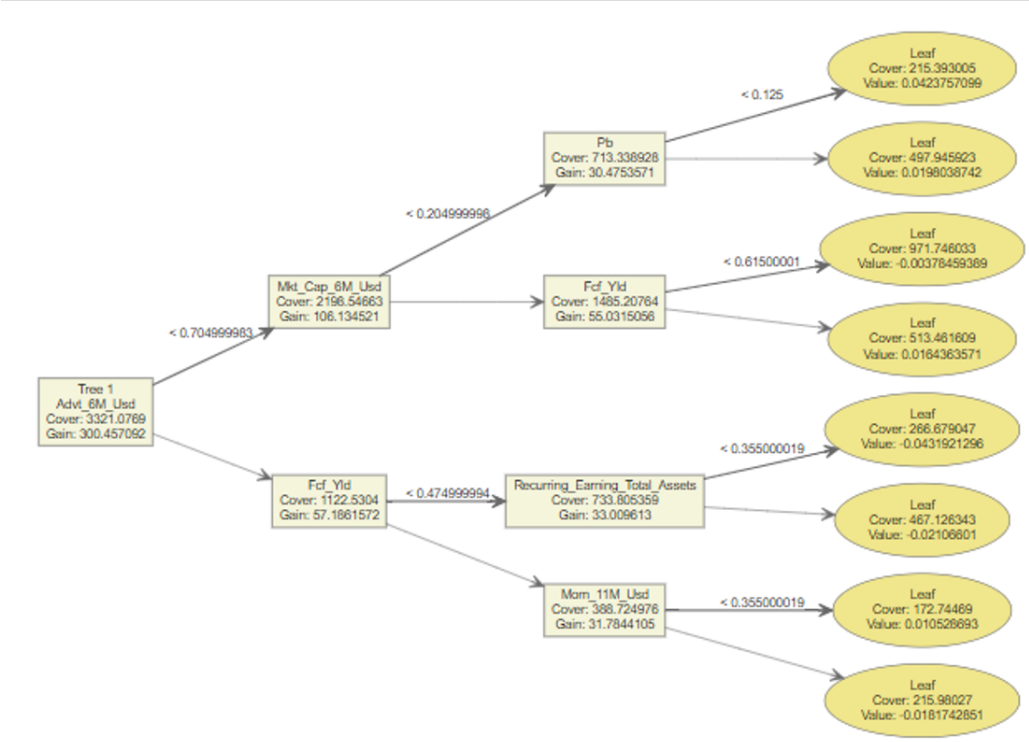

Figure 3. A partial picture of an example XGBoost tree

To find the best-performing model, we enable hyper-parameter tuning and let the gradient descent algorithm iterate over various parameter combinations and pick a model based on the AUC (Area Under the Curve) measure based on the ROC (Receiver Operating Characteristic) curve. For our given sub-sample, the AUC numbers as seen in Figure 4.

We see that the *Test* AUC tapers off quickly while *Train* fit is improved with iterations. This is not surprising since the useful information content of the data is limited in a similar fashion to the experiments conducted by [23].

The selected final model shows an ROC curve in Figure 5. The predictive ability looks poor but in the financial markets context, marginal improvements in the probability of picking Winners versus Losers may

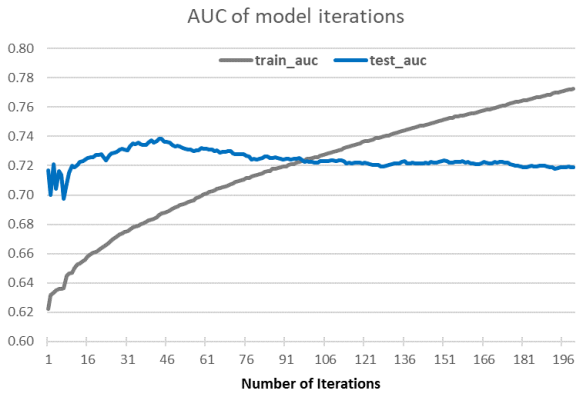

Figure 4. AUC of Test and Train over 200 Iterations

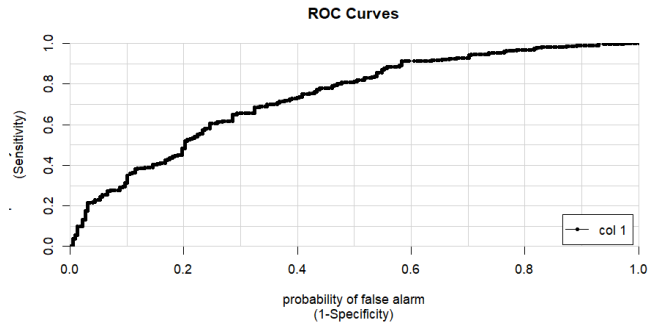

Figure 5. ROC Curve suggests very limited predictive potential but it is better than expected for stock returns

have significant economic meaning. The idea is not to reach high accuracy, such as in the X-ray image recognition problems, but to raise the odds somewhat, even if it is small.

Imagine running a hedge fund managing $10 billion, a 1% increase in the odds may amount to non-negligible gains. Therefore, in the context of stock returns, the results look interesting, and even remarkable. When we carry the model to a *Validation* sample that is later than the *Test* sample, we see that a similar outcome occurs.

Table 1 below presents the results of the Logit regressions of Predicted Probability on the Realised Probability of selecting Winners. Both the *Test* sample and Validation sample results confirm that the model-estimated probabilities have a statistically significant correlation with the actual outcomes. That is quite encouraging.

The *Test* sample used to produce the results is in 2002. If we were in 2003 now, and we had run the same method to get these results in 2003, would we recommend the ML-driven stock selection strategy as a useful model? Maybe, if we believed that the results are repeatable in the future. However, we did not produce any evidence on how repeatable the results could be.

At this point, it is important to remember the discussion about the 'instances of statistical results' being published, and sometimes commercially implemented. In the examples discussed earlier, an in many other similar work, the researchers report the incidents of interesting results appearing in their data sample, but do not proceed with further discussions on reliability or future usefulness. They conclude their work with the reporting of the statistical instances without analysing how easy it might be to find a fluke with the data and the ML machinery in hand. Those results do not reflect any 'learning' or 'AI', just like the results shown here so far do not.

Now, let's develop our example further by utilising more of the data sample. How would the results look if we were to re-run the modeling exercise over other periods, and then look at the performance of the portfolios that might have been constructed with the help of the ML-driven models?

Sample: Test. Logit Regression

Dependent Variable: Winner or Loser: 1 or 0

	Estimate	Std.Error	z-value
Intercept	-5.4	0.7	-7.7 ***
Predicted Odds of Being Winner	10.6	1.36	-7.8 ***

**** Significant at 0.001 level*

Sample: Validation Logit Regression

Dependent Variable: Winner or Loser: 1 or 0

	Estimate	Std.Error	z-value
Intercept	-5.8	0.4	14 ***
Predicted Odds of Being Winner	11.3	0.8	14.2 ***

**** Significant at 0.001 level*

Table 1: Do the predicted odds actually help predict the Winners?

When we repeat the exercise over different, consecutive samples and show that we are able to establish a relation between the odds predicted by the ML-driven model and the real odds of catching the Winners, we might have an 'AI-powered' strategy for stock investing.

It is common practice to apply a moving-window sampling to partition the time series data into *Train* and *Test* sub-samples so that the chronological consistency is maintained in the process. Randomized sampling over time does not work with time-series data due to the risk of look-ahead bias. Especially in the investment strategy development practices, researchers run the model-driven portfolio decisions over time with moving samples to demonstrate how the portfolios could have performed if the same decision rules or modeling methods were applied. It is called back-testing. Many academic publications also use the same procedure to validate their predictive modeling. (See Kelly et al. (2022) and Harvey et al. (2019) [19, 25] for a couple of examples.)

To see whether our ML-driven portfolio decision rule could work over time, let's repeat the XGBoost model fitting exercise over consecutive moving samples and construct portfolios according to the predicted odds of catching Winner stocks. As mentioned earlier, the objective is not to make highly accurate predictions of stock returns but to improve the odds for our bets in the gamble. At a point in time, we bet on roughly 150-200 stocks to buy (to go Long in finance lingo) and about the same number of stocks to sell (to go Short in finance lingo) out of about 1000 stocks. Among all those bets, if we can catch a few good ones, and avoid the bad outcomes each time, we can accumulate profits as we repeat the same process over and over.

We let our XGBoost model train over 36-month periods, as shown in Figure 9, then predict the Winner stocks in the consecutive 3-month Test period which is separated from the *Train* sample by +3-month gap to avoid any information leakage. We construct equal-weighted portfolios of stocks that are predicted to be likely to deliver Winner performance (i.e. top 80-percentile in that particular 3-month period) and we build another portfolio with the stocks that are predicted to be the least likely Winners. We calculate the return difference between the predicted Winner and Loser stock baskets for the period up to 2008. The accumulated return trajectory looks like the one shown in Figure 10.

The performance chart looks encouraging again. The AI machinery seems to be able to find a way to improve the odds of our 3-monthly bets on stocks. The evidence on the usefulness of the ML algorithms to guide the future stock return forecasts is accumulating, or it seems so.

Such cumulative return charts of back-tested portfolios are used widely as a historical validation tool in finance. Although it is helpful to run such experiments on historical data sets, the resultant performance charts may not reveal much about model reliability. In our case, for example, where we choose roughly 200

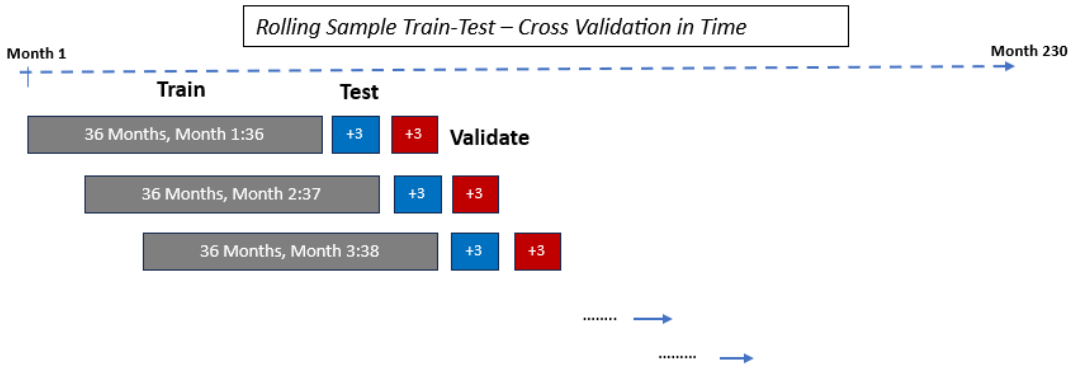

Figure 6. Model fitting with moving samples in time

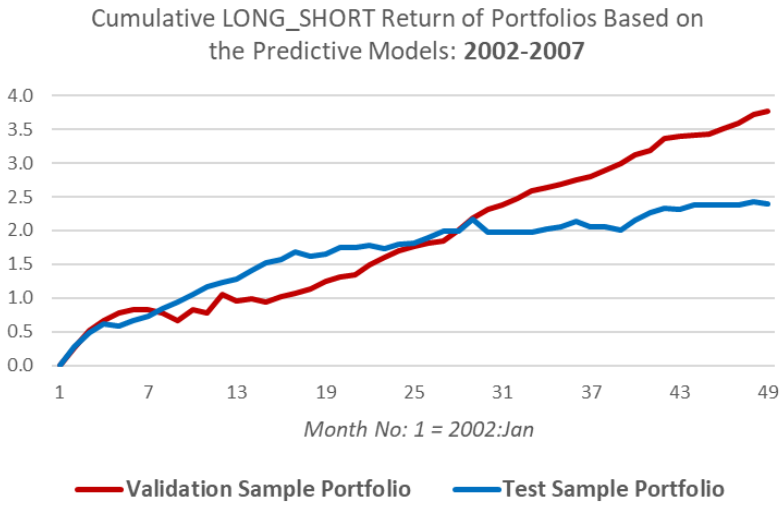

Figure 7. ML-driven model seems to deliver remarkable portfolio performance!

stocks among 1000, we have to acknowledge that there are countless (practically, pretty much infinitely many) portfolio combinations that may be shown to outperform another. It is highly unlikely not to randomly find a lucky portfolio among co many possible combinations.

On the other hand, if we divide our sample into much smaller sub-samples, 100 stocks among the available 1000 to fit our model, for example, the advantage of exploiting large data sets with ML algorithms fade away. Therefore, when we see back-testing exercises that are driven from ML models trained and tested on large data sets, we need to look into the drivers of results carefully to answer the following question: *Is the cumulative performance driven by a coincidental sequence of luck or by the accurate predictions of the model?*

In commercial applications, such as the AI-powered ETF products mention earlier, the questions about the probable sequential luck in their back-tests are completely omitted. Such an inquiry is against the commercial incentives to ride the AI wave of our time. Additionally, academics also tend to rely heavily on back-test results to show some evidence of validation for their models. Those practices are criticized in a growing number of papers such as [13], [14] and [25].

Now, let's make an attempt to shed some light onto the likelihood of 'sequential luck' in our case. We see that the ML-based model is able to help us accumulate positive returns with the historical sample prior to 2008. Are those positive returns driven by the models' successful predictions or are we picking up some lucky draws generated by the complex decision tree models?

In order to answer that question, we can run Logit regressions just like the ones presented earlier. If the 'predicted odds of being a Winner stock' correlates with 'actually being one of the Winner stocks' consistently over sequential samples, then we can build more confidence on the reliability of the data and the methods

employed.

In Figure 8, ideally, we would like to observe the z-values pile up in the second quadrant, in and around the blue shaded area. We see that the dots are slightly tilted towards that area, but it is hard to argue for a significant cluster. Actually, if we remove 2-3 outliers from the picture, the chart becomes an evenly spread out scatter centered around zero. That suggests that some luck is involved in upward-trending back-tests.

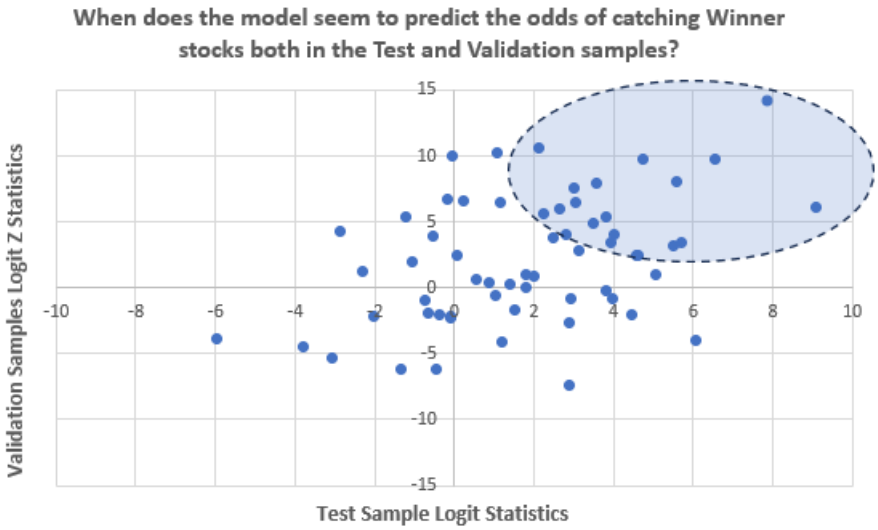

Figure 8. Z-value of Logit regressions of model prediction on real outcomes - Test vs. Validation samples

Random luck should converge to an average of zero success in the long run. You might have some lucky streak from time to time, but it tends to correct over time. When we extend our sample further into the following 10 years, we see that outcome.

In Figure 9, the *Test* sample continues to accumulate some positive return since, during the hyper-parameter tuning and model-selection process, the iterative algorithm uses the Test sample to optimize accuracy. However, when we try to implement that 'optimal model' in the following validation period, we see that the model does not bring any value.

If we were in 2009, for example, and got excited with the back-test results of our smart, AI-powered setup and implemented it as an investment strategy, we would end-up losing great sums of money- just like many other similar strategies do all the time.

The simple case discussed above clearly demonstrates the importance of collecting as many instances of statistical results as possible to gauge the reliability of the models fitted to historical samples. Unfortunately, neither the financial industry nor the academic researchers seem to have the necessary focus on reliability due to the ongoing rush to produce the next interesting statistical machinery that seems to show an instance of predictive success. Many end up reporting their lucky draw with an 'outcome bias'.

Forecasting is not only about predictive accuracy but also about estimation of the model risk. Machine learning models that are over-fitted onto the single sequence of observed history carry substantial reliability risks. However, the trendy labels with suggestive words such as 'learning' and 'intelligence' seem to create some illusion about the models' limited capabilities especially with historical data in social contexts. The luck factor and observation bias hiding behind the complex algorithms is a much bigger problem than it is usually discussed.

3 Conclusion

'Learning' is essentially about figuring things out with experience. AI technologies allow the computers to gain and simulate experience by using large amounts of data. As long as we can define the objective and formulate the related optimization problem, iterations over patterns in large data sets help us distill the

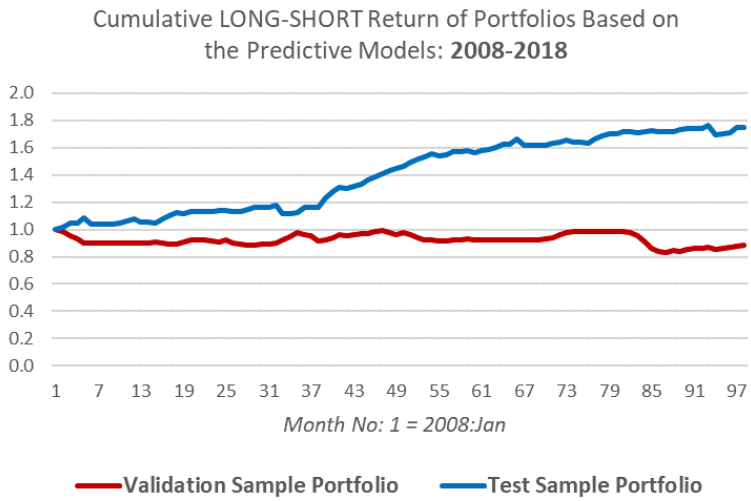

Figure 9. Performance in the post-2008 sample converge to zero

mechanisms that bring practical solutions. That is what we observe in many fields from automation to GO playing. The Boston Dynamics robot dog, Spot, for example, learns to walk over obstacles after processing the data of all previous falls to make improvements on its target.

A historical data set generally does not provide much information on repeated failures. It shows the outcomes captured in a certain set of circumstances that may radically change over time. To adopt AI techniques with such data sets might help us uncover some patterns that occurred in the past, but it would not necessarily yield any reliable predictions for the future.

The researchers in social fields, such as economics and finance, generally assume that an out-of-sample back-test can be used as an evidence of reliability. They present their results without discussing how likely it is to find such a back-test result simply by luck. In some cases, like the one developed in this paper, it may be very easy to find just by construct. The way we design the prediction target, the data we use, and the computational tools we implement might become a powerful combination to produce many statistical flukes. Therefore, while designing forecast models, the researchers need to extend their results into a detailed discussion on reliability.

Given that we are able to discover statistical patterns and validate them with the historical data, how useful should we expect those findings to be in the future?

To answer that question, we need to run empirical experiments to show what the results could have been if we had actually implemented similar models discovered in the past. The analysis and discussions in this paper offer some practical approaches to design such experiments. The results show that the incidental statistical discoveries may crumble easily, no matter how they may look convincing in the past. The finance literature is full of such examples.

As AI-powered applications proliferate many fields in business and academia, it is important to acknowledge the rising risk of statistical deception as a byproduct of careless and lazy model implementations. Extra care and regulation may be needed in the areas where artificial intelligence blends with too much natural human ignorance.

4 Discussion

The complexities related to the implementation of predictive machinery in financial investment and economic policy are actually go far beyond the reliability issues. For example, if large asset managers start implementing investment decision rules based on similar models, and if those models start to trigger correlated decision signals, they might generate self-fulfilling fluctuations in the market. Not only the actions become more predictable but also the models might induce cascades of decisions that are chasing each other.

Cascading actions are a common phenomenon in the financial markets. The AI models carry the risk of amplifying the cascades with automated herding behavior. Then the models drive their own validation success, and demise.

The nature of the problem is quite complex. In social settings, if enough people believe in something, that belief actually becomes the truth. If all believe that AI has huge economic potential and NVIDIA will be the company to benefit from that economic potential, and then invest in the company, NVIDIA stock price surges, pushes down cost of capital and triggers more investment decisions by the company, in a circular manner. AI-driven decision rules can work to build that self-fulfilling circuit. Therefore, the reliability and usefulness the models become rather fluid and stochastic.

An analogy from image recognition models would be as follows: In social settings and markets, if there are enough number of people that believes in a model that says the image is a cat, the image actually becomes a cat, no matter what it was to start with.

Such complexities will probably be the subject of other papers.

Acknowledgements

I thank Tony Guida for his valuable discussion in the AIFI - AI in Finance workshop, and for making the data sets available.

References

[1] Berman R, Sweeney P, Katari S. Machine learning for corporate strategy: An application to strategic decision-making in the pharmaceutical industry. Strategic Management Journal. 2021;42(7):1215-37.

[2] Lee J, Chen Z. Predictive analytics for employee success and retention: A machine learning approach. Journal of Business Research. 2020;116:372-80.

[3] Chen T, Guestrin C. Predicting economic and political stability using big data and machine learning. Political Analysis. 2016;24(3):293-311.

[4] Erel I, Stern LH, Tan C, Weisbach MS. Selecting Directors Using Machine Learning. The Review of Financial Studies. 2021;34(7):3226-64. Available from: https://doi.org/10.1093/rfs/hhaa133.

[5] Ross SA. The Arbitrage Theory of Capital Asset Pricing. Journal of Economic Theory. 1976;13(3):341-60.

[6] Merton RC. An Intertemporal Capital Asset Pricing Model. Econometrica. 1973;41(5):867-87.

[7] Fama EF, French KR. The Cross-Section of Expected Stock Returns. The Journal of Finance. 1992;47(2):427-65.

[8] Fama EF, French KR. Common Risk Factors in the Returns on Stocks and Bonds. Journal of Financial Economics. 1993;33(1):3-56.

[9] Cochrane JH. Presidential Address: Discount Rates. The Journal of Finance. 2011;66(4):1047-108.

[10] Carhart MM. On Persistence in Mutual Fund Performance. The Journal of Finance. 1997;52(1):57-82.

[11] Fama EF, French KR, Carhart MM. Characteristics, Covariances, and Average Returns: 1929 to 1997. The Journal of Finance. 2000;55(1):389-406.

[12] Fama EF, French KR. A Five-Factor Asset Pricing Model. Journal of Financial Economics. 2015;116(1):1-22.

[13] López de Prado M, Bailey DH. Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance. Notices of the American Mathematical Society. 2014;61(5):458-71.

[14] Bailey DH, López de Prado M. Backtest Overfitting. Journal of Computational Finance. 2014;18(2):21-36.

[15] López de Prado M. Why has Factor Investing Failed?: The Role of Specification Errors. SSRN Electronic Journal. 2020.

[16] Harvey CR, co authors. ... and the Cross Section of Returns. Review of Financial Studies. 2016;29(3):580-637.

[17] Harvey CR, Liu Y, Zhu H. Lucky Factors. SSRN Electronic Journal. 2016.

[18] Kelly B, co authors. The Virtue of Complexity in Return Prediction. SSRN Electronic Journal. 2022.

[19] Kelly B, co authors. Factor Models, Machine Learning, and Asset Pricing. NBER Working Paper Series. 2022;(w30599).

[20] Chen AY, Zimmermann T. Open Source Cross-Sectional Asset Pricing. Journal of Finance. 2020;75(3):1539-86.

[21] Chen AY, Velikov M. Zeroing in on the Expected Returns of Anomalies. Journal of Financial Economics. 2021;142(2):679-703.

[22] Belkin M. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. arXiv preprint arXiv:210514368. 2021. Available at arXiv: https://arxiv.org/abs/2105.14368.

[23] Noguer i Alonso M, Srivastava S. The Shape of Performance Curve in Financial Time Series. SSRN Electronic Journal. 2023.

[24] Guida T. Machine Learning for Factor Investing: R Version. Hoboken, NJ: Wiley; 2020.

[25] Harvey CR, co authors. Machine Learning in Finance: The Case of Missing Factors and Alternative Anomalies. SSRN Electronic Journal. 2019.

Appendix A: Data set

The data set on stock returns and attributes is a courtesy of the work by Guida (2020) [24]. The variable descriptions and exploratory data analysis can be found at: https://www.mlfactor.com/data-description.html

All feature variables are scaled and normalised. The details are not included to keep this document in manageable length.

The R codes used in this paper are available by request from the author.