All that glitters is not gold: Comparing backtest and out-of-sample performance on a large cohort of trading algorithms

Authors: Dr. Thomas Wiecki, Andrew Campbell, Justin Lent, Dr. Jessica Stauth
Affiliation: Quantopian Inc

Abstract

When automated trading strategies are developed and evaluated using backtests on historical pricing data, there exists a tendency to overfit to the past. Using a unique dataset of 888 algorithmic trading strategies developed and backtested on the Quantopian platform with at least 6 months of out-of-sample performance, we study the prevalence and impact of backtest overfitting. Specifically, we find that commonly reported backtest evaluation metrics like the Sharpe ratio offer little value in predicting out of sample performance ($R^2 < 0.025$). In contrast, higher order moments, like volatility and maximum drawdown, as well as portfolio construction features, like hedging, show significant predictive value of relevance to quantitative finance practitioners. Moreover, in line with prior theoretical considerations, we find empirical evidence of overfitting – the more backtesting a quant has done for a strategy, the larger the discrepancy between backtest and out-of-sample performance. Finally, we show that by training non-linear machine learning classifiers on a variety of features that describe backtest behavior, out-of-sample performance can be predicted at a much higher accuracy ($R^2 = 0.17$) on hold-out data compared to using linear, univariate features. A portfolio constructed on predictions on hold-out data performed significantly better out-of-sample than one constructed from algorithms with the highest backtest Sharpe ratios.

Introduction

When developing automated trading strategies, it is common practice to test algorithms on historical data, a procedure known as backtesting. Backtest results are often used as a proxy for the expected future performance of a strategy. Thus, in an effort to optimize expected out-of-sample (OOS) performance, quants often spend considerable time tuning algorithm parameters to produce optimal backtest performance on in-sample (IS) data. Several authors have pointed out how this practice of backtest "overfitting" can lead to strategies that leverage to specific noise patterns in the historical data rather than the signal that was meant to be exploited (Lopez de Prado [2013], Bailey et al. [2014a]; Bailey et al. [2014b]). When deployed into out-of-sample trading, the expected returns of overfit strategies have been hypothesized to be random at best and consistently negative at worst.

The question of how predictive a backtest is of future performance is as critical as it is ubiquitous to quantitative asset managers who often, at least partly, rely on backtest performance in their hiring

and allocation decisions. In order to quantify backtest and out-of-sample performance, a large number of performance metrics have been proposed. While the Sharpe ratio (Sharpe [1966]) is the most widely known, it is probably also the most widely criticized (Spurgin [2001], Lin, Chou [2003]; Lo [2009]; Bailey, Lopez de Prado [2014]) . A large number of supposedly improved metrics, such as the information ratio or Calmar ratio (Young [1991]), have been proposed, but it is unclear what predictive value each metric carries.

Backtest overfitting also appears to be a problem in the academic literature on quantitative finance where trading strategies with impressive backtest performance are frequently published which do not seem to match their OOS performance (for studies on overfitting see e.g. Schorfheide and Wolpin [2012]; McClean and Pontiff [2012]; Lopez de Prado [2013]; Bailey [2014a]; Bailey [2014b]; Beaudan [2013]; Burns [2006]; Harvey et al. [2014]; Harvey, Liu, & Zhu [2016]). A recent simulation study by Bailey et al. [2013] demonstrates how easy it is to achieve stellar backtest performance on a strategy that in reality has no edge in the market. Specifically, the authors simulate return paths with an expected Sharpe Ratio of 0 and derive probabilities to achieve Sharpe Ratios well above 1 after trying a few strategy variations under a limited backtest time-frame. When no compensatory effects are present in the market, selecting such a strategy based on in-sample Sharpe Ratio will lead to a disappointing out-of-sample Sharpe Ratio of 0. However, when assuming such compensatory market forces like overcrowded investment opportunities to be at play, selecting strategies with high in-sample Sharpe ratio would even lead to negative out-of-sample Sharpe ratio. As these results are purely theoretical, it is not clear which of these two relationships – zero or negative correlation – between IS and OOS performance exist in reality.

In this study, we aim to provide an empirical answer to the relationship between the IS and OOS performance based on data set and compare various performance metrics that have been proposed in the literature. Towards this goal, we have assembled a data set of 888 unique US equities trading algorithms developed on the Quantopian platform and backtested from 2010 through 2015 with at least 6 to 12 months of true OOS performance. Quantopian provides a web-based platform to research, develop, backtest and deploy trading algorithms. To date, users of the platform have run over 800,000 backtests. While the site terms-of-use strictly prohibit direct investigation of algorithm source code, we are granted access to detailed data exhaust, returns, positions, and transactions an algorithm generates when backtested over arbitrary date ranges. As the encrypted algorithm code is time-stamped in our database, we can easily determine exactly what historical market data the author had access to during development. We call this time prior to algorithm deployment the in-sample period. The simulated performance accumulated since an algorithm's deployment date represents true out-of-sample data.

As we will show below, backtest performance of single metrics have very weak correlations with their out-of-sample equivalent (with some exceptions). This result by itself might lead to the conclusion that backtests carry very little predictive information about future performance. However, by applying machine learning algorithms on a variety of features designed to describe algorithm behavior we show that OOS performance can indeed be predicted.

Methods

Data set

Our initial sample consists of 7152 algorithms developed and backtested on the Quantopian platform (https://www.quantopian.com). The algorithms in our sample represent a wide range of strategy styles (technical, fundamental, mean-reversion, momentum, etc.), trading frequencies (buy-and-hold/intraday), portfolio sizes (1 - 500 stocks), and portfolio structures (long-only/short-only/dollar-neutral). This heterogeneity adds valuable variance to our IS/OOS feature dataset. We applied several filtering steps to remove duplicates, outliers, and algorithms that likely do not represent meaningful attempts at a profitable strategy. Specifically, we removed strategies that only trade a single stock, or had a Sharpe ratio less than -1.0, or were backtested over less than 500 days in total, or were not invested in the market on at least 80% of trading days. We also removed outlier strategies where a feature (see below for information on how features are computed) deviated more than 4 standard deviations from its population mean. Finally, we had to remove duplicate strategies which are frequent because of a “cloning” feature on Quantopian that allows quants to copy a strategy that was shared on the community forums. To filter these clones, we removed algorithms that traded the exact same number of trading days, maximum, median and minimum stocks or were correlated more 0.95 with strategies already in our sample. Ultimately, these preprocessing steps removed a large part of our initial data set and left us with 888 algorithms. We experimented with various preprocessing parameters and filters and decided on these as they remove many non-representative or duplicate strategies but still leave us with enough data to perform meaningful analyses. Results reported below overall were robust to specific preprocessing choices.

Each algorithm in our sample pool was backtested from 2010 through the end of 2015 using the open-source Zipline trading simulator (https://zipline.io). Zipline simulates split-adjustment, transaction costs, order delays, liquidity constraints, as well as market impact and slippage. For all simulations, minute-bars trade level pricing data were used. Since all the algorithms in our sample were written between January and June in 2015, each backtest contains a minimum of 6 months of OOS performance data. It is important to emphasize that all algorithm code is versioned in a point-in-time database at time of creation, the algorithm's logic could not have been updated or adjusted at any point during this OOS period. From these backtests, we utilize simulated transactions, end-of-day positions, and daily returns to compute risk metrics and features. We also compiled platform usage data on the total number of backtest days each algorithm was tested over by its author prior to deployment. Our hypothesis is that this backtesting activity metadata will capture the impact of strategy development technique on OOS performance.

The main analyses below are carried out using frequentist linear regression methods. To gain more confidence into our results we have also performed Bayesian linear regressions which resulted in similar results but are omitted due to the widespread familiarity with frequentist statistics.

Feature extraction

As the raw returns, positions, and transactions data that constitute a backtest are often difficult to compare across algorithms, the extraction of universally applicable features is a critical step in any cross-sectional comparison of strategy performance. We constructed features based on point estimates of several well-known performance and risk metrics such as the Sharpe ratio, information ratio, Calmar ratio, alpha, beta, maximum drawdown, as well as annual returns and volatility. In addition, we included as features metrics that track algorithm behavior based on returns (including skew, kurtosis, standard deviation of rolling beta with a 6-month window), positions (including median and maximum position concentration, total number of tickers traded), and transactions (including average % daily turnover, percent of winning trades). In an effort to capture the influence of strategy development technique and potential overfitting, we have also included a feature for the total number of backtest days each algorithm was tested over prior to deployment.

As risk metrics are not necessarily stable over the course of an algorithm's full backtest, we attempt to capture higher moments of the returns time series by using rolling risk metrics and structural components as additional features. For example, we computed a feature from the standard deviation of 6-month rolling Sharpe ratio over the IS/OOS periods.

Our feature extraction methodology relied heavily on the open-source performance analysis library Pyfolio (https://www.github.com/quantopian/pyfolio). In total, we constructed 57 individual features applied to IS and OOS data separately. For a list of all features and their computation, we refer to the supplement.

Machine Learning

As we will show below, linear methods did not show high predictability of individual backtest performance measures on OOS profitability. We next asked if non-linear regression methods trained on the full feature set could do better at predict OOS Sharpe ratio. Towards this goal, we explored a number of machine learning techniques, including Random Forest and Gradient Boosting, to predict OOS Sharpe. To avoid overfitting, all experiments used a 5-fold cross-validation during hyperparameter optimization and a 20% hold-out set to evaluate performance. In addition to training our own classifiers, we also utilized the DataRobot platform (https://www.datarobot.com) to test a large number of preprocessing, imputation and classifier combinations.

Consistent with prior literature on the application of machine learning to investing (Martinez et. al [2009]; Kearns and Nevmyvaka [2013]), we attempt to measure the performance of our classifier as a portfolio selection method. To simulate selection performance, we created an equal-weighted portfolio of the top 10 algorithms from the hold-out set based on their predicted OOS Sharpe ratio. Since the assets in our simulated portfolio are algorithms that can be backtested with slippage, commissions and latency assumptions, we avoid many of the pitfalls that have hindered previous studies' simulation efforts. We contrast the resulting portfolio returns to those of random selection (Burns [2006]) and selection by highest IS Sharpe ratio.

Results

In-sample vs out-of-sample comparison

Our first set of analyses aims to evaluate the degree to which various IS performance metrics correlate with the same metric computed over only the OOS period. As metrics like the Sharpe ratio are noisy measures themselves (Lo [2002]) we first tested whether correlations hold across IS periods. We thus compared the Sharpe ratio from the last year of each algorithm's IS period to that of the preceding IS period. Regression analysis revealed a strong and highly significant linear relationships (Pearson $R^2 = 0.21$; $p < 0.0001$) which establishes a rough baseline as to the maximum effect size to expect when comparing performance metrics across IS and OOS periods.

When comparing IS and OOS periods, we found a weakly negative but highly significant correlation between annual returns periods (Pearson $R^2=0.015$; $p < 0.001$; figure 1a). Contrary, Sharpe ratio IS was positively correlated with Sharpe ratio OOS (Pearson $R^2=0.02$; $p < 0.0001$; figure 1b). Interestingly, using only the last backtest year to compute IS Sharpe ratio increased predictability of OOS Sharpe ratio suggesting a recency effect (Pearson $R^2=0.05$; $p < 0.0001$; not shown). Next, Sortino ratio IS also showed weak predictability of Sortino ratio OOS (Pearson $R^2=0.02$; $p < 0.0001$; figure 1d). All other IS vs OOS performance metrics were not significant with Pearson R^2 values below 0.005, including information ratio (figure 1c), Calmar ratio (figure 1e), and financial alpha (figure 1f).

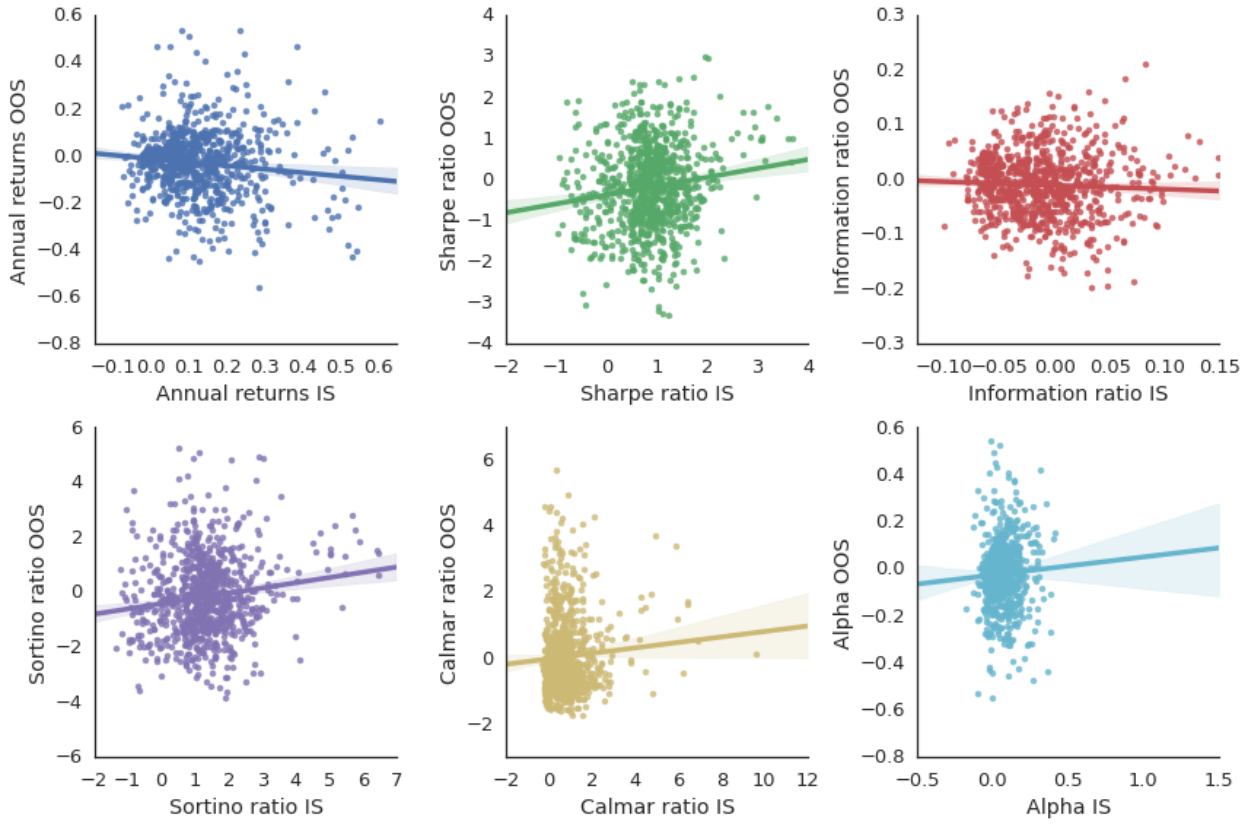

Figure 1: Scatter plots of in-sample (IS) vs out-of-sample (OOS) values of (a) annual returns, (b) Sharpe ratio, (c) information ratio, (d) Sortino ratio, (e) Calmar ratio, and (f) financial alpha. The line indicates the best fitting linear regression with the shaded area showing 5% and 95% confidence intervals. None of the performance metrics show a significant correlation between their IS and OOS periods.

It is curious to observe a negative correlation between IS and OOS annual returns but a (slightly) positive relationship between IS and OOS Sharpe ratio, which has average returns in its nominator. We propose this pattern can be explained by two findings (figure 2): (i) a positive correlation between mean returns and annual volatility ($p < 0.001$), in line with Kakushadze & Tuchinsky [2015] and Kakushadze, Lauprete and Tulchinsky [2015], and, (ii) a significant interaction between annual volatility and annual returns on OOS Sharpe ratio ($p = 0.009$). In other words, IS Sharpe ratio can be increased either by increasing mean returns or decreasing volatility. It appears that the former is more prone to overfitting. This could suggest that some quants in our data set focused on maximizing returns while not taking volatility into account.

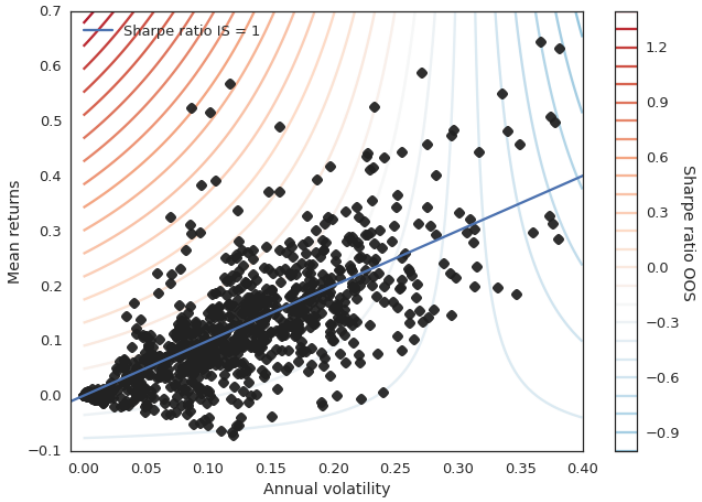

Figure 2: Scatter-plot of annual volatility (IS) vs mean returns (IS) with interaction gradient of OOS Sharpe ratio. The blue line through the origin represents an IS Sharpe ratio of 1. The area above the “Sharpe ratio IS = 1”-line represents Sharpe ratios > 1, while the area below Sharpe ratios < 1. As can be seen, there is a strong linear relationship between annual volatility and mean returns with an interaction of OOS Sharpe ratio. Thus, strategies that increase their Sharpe ratio by taking on excessive volatility have worse OOS Sharpe ratio than those that keep volatility low.

Interestingly, tail-ratio (i.e. the ratio between the 95th and 5th percentile of the returns distribution) showed a stronger significant correlation with OOS Sharpe ratio than IS Sharpe ratio did (Pearson $R^2 = 0.025$; $p < 0.0001$). Moreover, risk metrics that aim to quantify volatility alone like annual volatility (Pearson $R^2 = 0.67$; $p < 0.0001$), and maximum drawdown (Pearson $R^2 = 0.34$; $p < 0.0001$) had statistically significant correlations between their IS and OOS period.

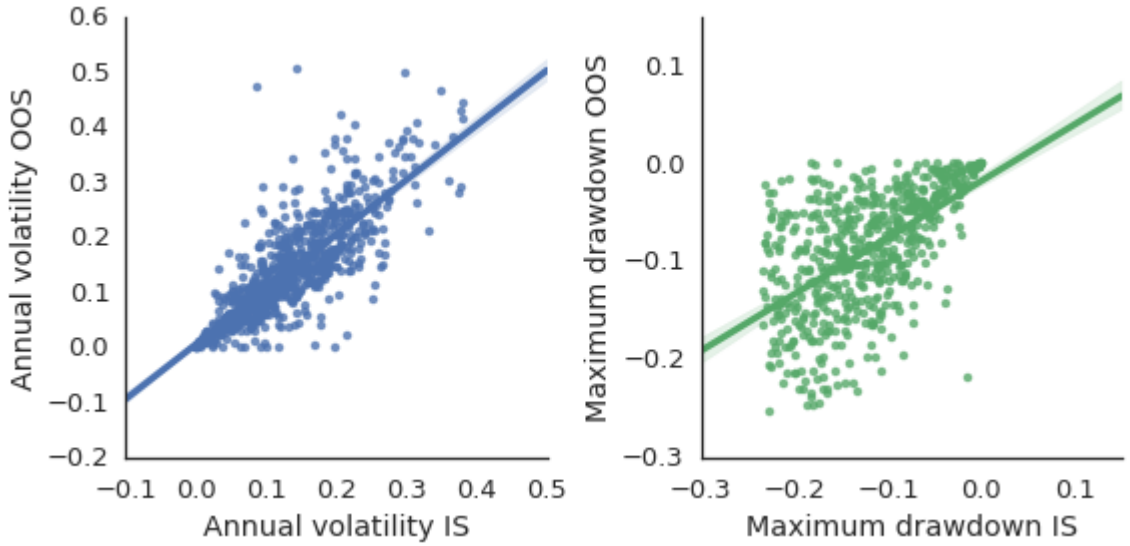

Figure 3: Scatter plots of in-sample (IS) vs out-of-sample (OOS) values of (a) annual volatility and (b) maximum drawdown. The line indicates the best fitting linear regression with the shaded area showing 5% and 95% confidence intervals.

Influence of number of backtests on Sharpe ratio

Several studies have shown that IS Sharpe ratio can be inflated by testing different configurations of a strategy (Bailey, Borwein & Lopez de Prado [2014]). As quants can run backtests repeatedly over varying time-ranges, we computed the total number of days a strategy was backtested over in the course of its development. As this variable is positive and was found to have a very long tail, we log-transformed it. To quantify the amount of overfitting, we computed Sharpe ratio shortfall (Sharpe ratio IS - Sharpe ratio OOS). In line with previous theoretical predictions, we confirmed a weak (Spearman R^2=0.017) but highly significant ($p < 0.0001$) positive correlation between the logarithm of the backtest days and Sharpe ratio shortfall (see figure 3). Interestingly, annual volatility IS also showed a positive correlation with Sharpe ratio shortfall ($R^2 = 0.02$; $p < 0.0001$) which corroborates our previous finding that backtests with higher volatility appear to suffer more from overfitting.

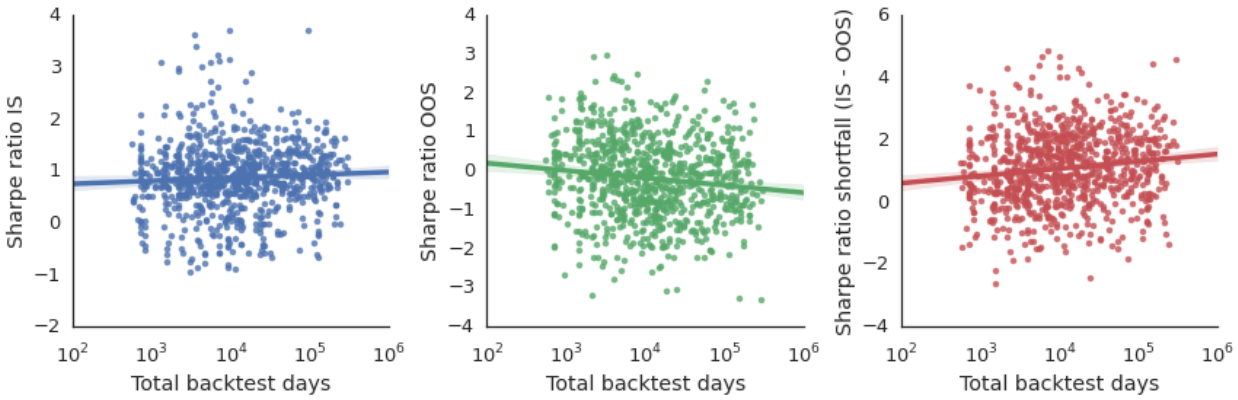

Figure 4: Scatter plots of total backtest days a quant ran for a strategy (in log-scale) vs (a) in-sample (IS) Sharpe ratio, (b) out-of-sample (OOS) Sharpe ratio, and (c) Sharpe ratio shortfall (i.e. IS Sharpe ratio - OOS Sharpe ratio). The line indicates the best fitting linear regression with the shaded area showing 5% and 95% confidence intervals. As can be seen, the more backtests a quant ran, the higher the IS Sharpe ratio, the lower the OOS Sharpe ratio, and the larger the shortfall.

Turn-over and hedging

In addition, we asked if certain parameters describing the behavior of a trading algorithms influence performance. In regards to backtest overfitting, we asked if strategies that made more independent bets, as measured by the monthly portfolio percent turnover, would be harder to overfit. A multilinear regression of the logarithm of monthly turnover and the logarithm of user backtest days onto Sharpe ratio shortfall failed to reject the null-hypothesis for turnover ($p=0.074$). These results, while surprising to us, are in line with Kakushadze and Tulchinsky [2015] and Kakushadze, Lauprete & Tulchinsky [2015]. We have also made a surprise discovery of a negative correlation between turnover and number of backtest days (Pearson $R^2=0.04$; $p < 0.0001$) which could indicate that strategies with less turnover require more backtesting in order to achieve consistent results.

In addition, we tested whether being hedged or market-neutral would decrease volatility of the strategy as market-perturbations would be expected to have a smaller impact on the strategy's returns. As many strategies in our sample had a minimum hedge-ratio of 0 (due to being long-only), we split strategies into buckets of a min-hedge-ratio < 0.5 and > 0.5 and compared annual volatility in these two sets. Indeed, hedged strategies had significantly lower volatility IS ($t=5.78$; $p < 0.0001$) and OOS ($t=4.62$; $p < 0.0001$).

Machine Learning

Accuracy

Best results were achieved using the DataRobot platform (https://www.datarobot.com) which tests a huge number of preprocessing, imputation and classifier combinations. The best performance was achieved by an Extra Trees Regressor (Geurts, Ernst, and Wehenkel [2006]) with an average cross-validation Pearson R^2 of 0.18 and a Pearson R^2 of 0.17 on 20% hold-out data suggesting weak overfitting. Unfortunately, results were not always identical across different different folds suggesting our data set might still be too small for a confident result.

The most important features as determined by the random forest regressor (Breiman [2001]) can be appreciated in figure 5.

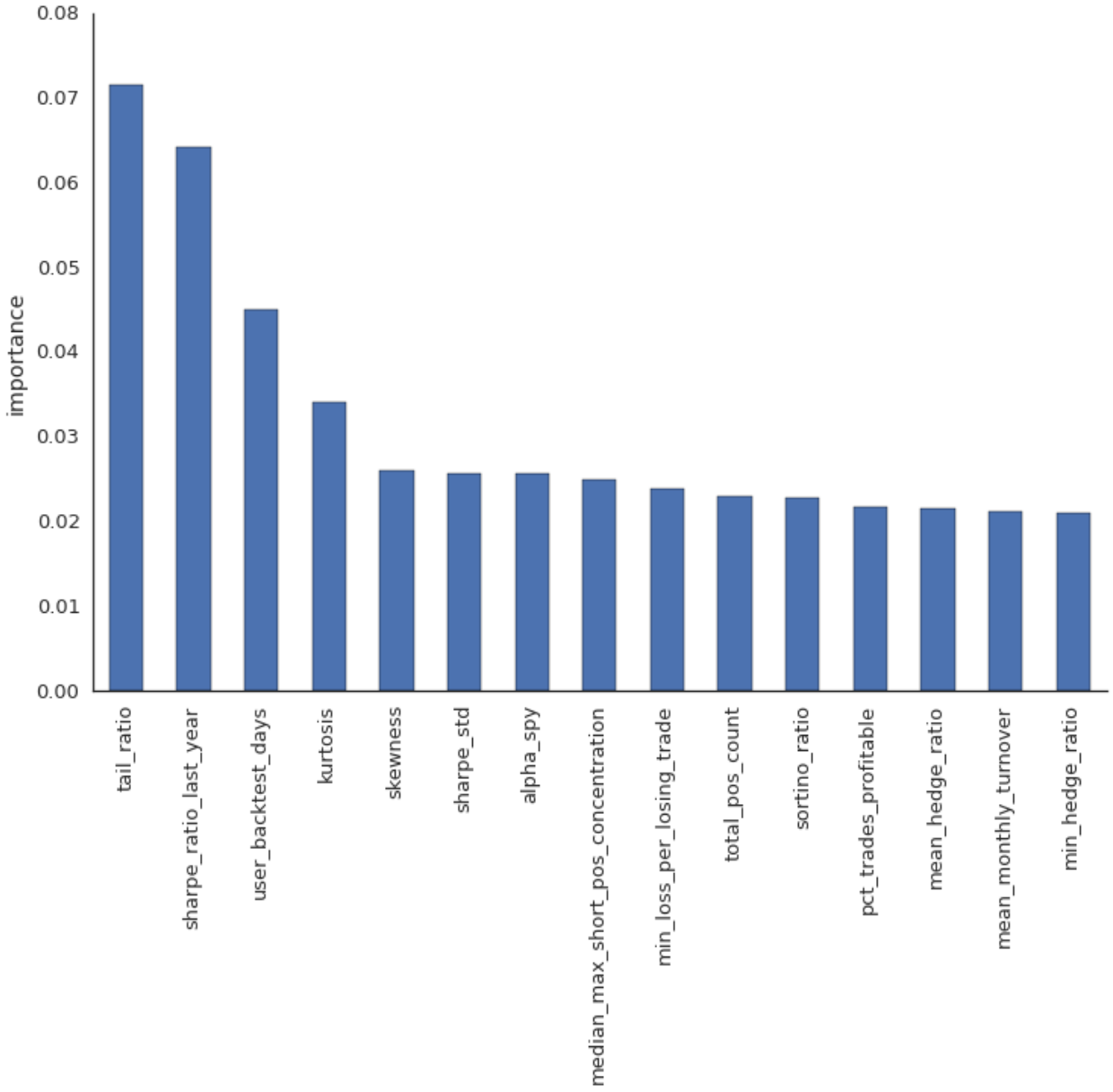

Figure 5: Top 15 most important features for predicting Sharpe ratio OOS as determined by a random forest regressor.

Performance of resulting portfolio

While the classical machine learning algorithm evaluation metrics described above suggest predictive significance for our non-linear classifier, the practical value of our machine learning methodology can only be evaluated by testing its profitability as a portfolio selection instrument. Towards this goal, we formed an equal-weighted portfolio out of 10 strategies with the highest Sharpe ratios as predicted by the Random Forest regressor on the hold-out set and computed their cumulative return (figure 6a) and the resulting Sharpe ratio. In addition, we compare this to 1000 random portfolios (Burns [2006]) of hold-out strategies and a portfolio formed by selecting strategies

with the 10 highest IS Sharpe ratios. We find that our ranking by predicted Sharpe portfolio performs better than 99% of randomly selected portfolios with a Sharpe ratio of 1.8 compared to the IS Sharpe ratio selection which proved better than 92.16% of random portfolios with a Sharpe ratio of 0.7 (figure 6b). Given the above result of weak predictability of IS Sharpe ratio it is surprising that it still performs reasonably well in this setting, albeit not at a statistically significant threshold compared to the random portfolios. These results do however show significant practical value for non-linear classification techniques compared to more traditional, univariate selection mechanisms when constructing portfolios of trading algorithms.

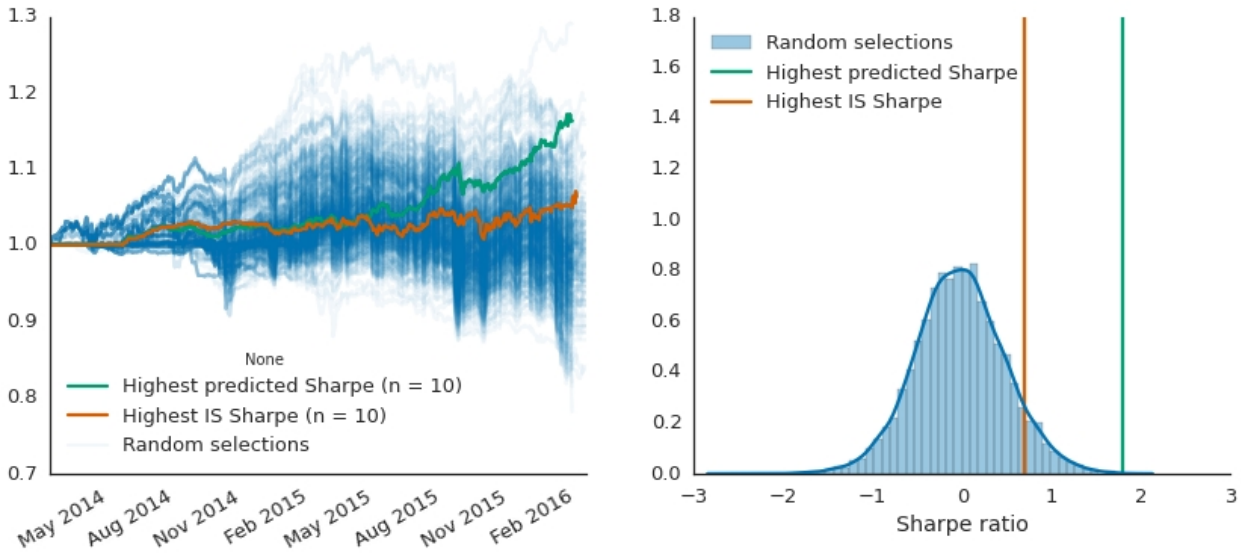

Figure 6: OOS performance of three portfolio selection mechanisms: (i) top 10 highest Sharpe ratios as selected by a random forest regressor, (ii) top 10 highest IS Sharpe ratios, and (iii) 1000 random selections of 10 algorithms each. a) cumulative OOS returns; b) Sharpe ratios of resulting portfolios.

Discussion

For the first time, to the best of our knowledge, we present empirical data that can be used to validate theoretical and anecdotal claims about the ubiquity of backtest overfitting and its impact on algorithm selection. This was possible by having access to a unique data set of 888 trading algorithms developed and tested by quants on the Quantopian platform. Analysis revealed several results relevant to the quantitative finance community at large – practitioners and academics alike.

Most strikingly, we find very weak correlations between IS and OOS performance in most common finance metrics including Sharpe ratio, information ratio, alpha. This result provides strong empirical support for the simulations carried out by Bailey et al. [2014]. More specifically, it supports the assumptions underlying their simulations without compensatory market forces to be present which would induce a negative correlation between IS and OOS Sharpe ratio. It is also interesting to compare different performance metrics in their predictability of OOS performance. Highest

predictability was achieved by using the Sharpe ratio computed over the last IS year. This feature was also picked up by the random forest classifier as the most predictive feature.

Comparing metrics computed over the full range shows that Sharpe and Sortino ratio have the highest OOS predictability, while the information ratio, alpha and Calmar ratio did not show significant effects. A closer analysis of the specific patterns across several IS and OOS performance metrics further revealed an interesting dynamic. While annual returns had a slightly negative correlation, Sharpe ratio, with mean returns in the nominator, showed a slightly positive correlation. This effect is explained by an interaction of mean returns and volatility on OOS Sharpe ratio. This finding is corroborated by a strong positive correlation between volatility and backtest overfitting. One potential explanation for this pattern is that quants were maximizing returns (but not Sharpe ratio) without considering the risk their strategy was taking on.

It is also interesting to consider performance and risk metrics that do not rely on average returns. Specifically, the tail-ratio was more predictive of OOS Sharpe ratio than IS Sharpe ratio itself. Of further note, volatility based risk metrics like standard deviation (annual volatility) and maximum drawdown hold very stable across IS and OOS periods.

Additionally, we find significant evidence that the more backtests a user ran, the bigger the difference between IS and OOS performance – a direct indication of the detrimental effect of backtest overfitting. This observed relationship is also consistent with Bailey et. al's [2014] prediction that increased backtesting of multiple strategy variations (parameter tuning) would increase overfitting. Thus, our results further support the notion that backtest overfitting is common and wide-spread. The observed significant positive relationship between amount of backtesting and Sharpe shortfall (IS Sharpe - OOS Sharpe) provides support for a Sharpe ratio penalized by the amount of backtesting (e.g. the "deflated Sharpe ratio" by Bailey & Lopez de Prado [2014]). An attempt to calibrate such a backtesting penalty based on observed data is a promising direction for future research.

Together, these sobering results suggest that a reported Sharpe ratio (or related measure) based on backtest results alone can not be expected to prevail in future market environments with any reasonable confidence. As we discussed above, this insight is gaining traction in the academic community (Harvey, Liu, & Zhu [2016]) and first examples exist of strategies being presented with a later follow-up of performance since initial publication. A notable performance difference between reported and post-published performance is also reported by McClean and Pontiff [2012] as well as Qu et al [2015]. These authors, however, attribute this mismatch not to overfitting but to market anomalies being arbitraged away after they become widely known. Under the assumptions that most strategies on Quantopian are not publicly known, this hypothesis is not sufficient to explain our results.

While the results described above are relevant by themselves, overall, predictability of OOS performance was low ($R^2 < 0.25$) suggesting that it is simply not possible to forecast profitability of a trading strategy based on its backtest data. However, we show that machine learning together with careful feature engineering can predict OOS performance far better than any of the individual measures alone. Using these predictions to construct a portfolio of strategies resulted in competitive

cumulative OOS returns with a Sharpe ratio of 1.2 that is better than most portfolios constructed by randomly selecting strategies. While it is difficult to extract an intuition about how the Random Forest is deriving predictions, we have provided some indication of which features it deems important. It is interesting to note that among the most important features are those that quantify higher-order moments including skew and tail-behavior of returns (tail-ratio and kurtosis). Together, these results suggest that predictive information can indeed be extracted from a backtest, just not in a linear and univariate way. It is important to note that we cannot yet claim that this specific selection mechanism will work well on future data as the machine learning algorithm might learn to predict which strategy type worked well over the specific OOS time-period most of our algorithms were tested on (for a more detailed discussion of this point, see the limitations section). However, if these results are reproducible on an independent data set or the strategies identified continue to outperform the broad cohort over a much longer time frame, it should be of high relevance to quantitative finance professionals who now have a more accurate and automatic tool to evaluate the merit of a trading algorithm. As such, we believe our work highlights the potential of a data scientific approach to quantitative portfolio construction as an alternative to discretionary capital allocation.

Limitations

Despite the robustness of our results, fairly high sample size, and agreement between our findings and the previous literature, several limitations could reduce the generality of our insights. Foremost, all algorithms were developed by Quantopian users. Demographic analysis of the Quantopian community reveals a heterogeneous and international group that range from engineers and academics with limited formal quant finance backgrounds to quant professionals with years of industry experience. Thus, the question arises whether our results extend to backtests developed exclusively by quantitative finance professionals who might employ methods that reduce overfitting when developing their trading strategies. We have made no attempt to identify these professionals within our data set and, therefore, we can not currently address that question. That said, verbal communications with practitioners who report similar patterns to those we have observed provides weak anecdotal evidence in favor of our study's representativeness of the professional investment industry.

Similarly, as we do not know the logic of the trading algorithms behind each observable backtest, it remains an open question how representative our results are of the academic literature describing compelling backtest performances of new trading algorithms. There are, however, individual cases where an algorithm author decided to share his or her implementation details with the Quantopian community. For example, the first winner of the first Quantopian Open trading competition revealed an implementation of a long-only mean-reversion algorithm suggested in a previously published study (Li & Hoi [2012]). While showing attractive IS performance, the algorithm lost money after being deployed with real capital. Limited OOS time as well as subtle implementation differences prohibit strong claims about a lack of validity of the research backing this particular strategy. However, it does show that our data set contains strategies published in the academic literature.

Another caveat to the generality of our results are the overlapping time periods over which the algorithms in our sample were developed and tested. As described above, the IS period for our pool covered a period from 2010 through June of 2015, a bull market with relatively little volatility, while our OOS period from June 2015 to February 2016 showed a flat-to-bear market with medium-to-high volatility. Thus, it is possible that the weak correlation between IS and OOS mean returns we observe is due to a shifting market regime and that our strategies would have continued to perform consistently with their backtests had there been no change in market regime. While we can not rule this effect out without access to more data from different market regimes, we have tested whether our observed patterns hold when our sample is limited to market-neutral strategies. Theoretically, a market-neutral strategy should not be as strongly affected by a change in market-regime. We thus selected strategies with a beta between -0.15 and 0.15 and confirmed that our reported results are reproduced in this subset. The overlap in time might also limit the generalizability of our ML results as the classifier could learn to predict which strategy type worked well over our specific OOS period, rather than any future time-period. Accumulating more OOS time should allow us to answer this question.

Finally, in this study, we focused primarily on OOS predictability based on backtests alone. However, many allocators, investors and quantitative hedge fund managers place great emphasis on existing OOS data when selecting strategies for deployment. In our experience with allocation decisions, we have adopted a requirement of least 6 months of OOS data to give special attention to any inconsistencies between IS and OOS performance. Future research will thus place a greater emphasis on the prediction of OOS performance based on backtest *and* OOS data. This could be achieved by splitting the existing OOS period in our data set.

Acknowledgements

We would like to thank the Quantopian community for producing the strategies that are the foundation of this analysis. We are also grateful to John Fawcett, CEO of Quantopian, for generously supporting this research and contributing to frequent discussions about our results. Moreover we would like to thank Jonathan Larkin for useful feedback on the manuscript. Finally, we would like to express our gratitude to the DataRobot team for providing us with access to their platform.

Appendix

Feature Name	Description
trading_days	number of days days from the algo's first trade to the OOS date
sharpe_ratio	mean(returns) / std(returns) * sqrt(252)
sharpe_ratio_last_year	Sharpe ratio over the last IS year

annret	annualized returns
annvol	annualized volatility of daily returns
skewness	skewness of daily returns distribution
kurtosis	kurtosis of daily returns distribution
stability	R-squared error of a linear fit to the cumulative log returns
beta_spy	Correlation between daily returns and daily returns of S&P 500 index returns
alpha_spy	Annualized returns in excess of returns resulting from correlation with the S&P 500 index
information_ratio	mean(returns - S&P500 returns) / std(returns - S&P500 returns)
beta_std	Standard deviation of rolling 6 month beta to the S&P 500
sharpe_std	Standard deviation of rolling 6 month Sharpe ratio
sortino_ratio	mean(returns) / std(returns[returns < 0])
drawdown_area	Annualized area of drawdown periods (bounded by high water mark and cumulative returns curve)
max_drawdown	Maximum peak to trough drawdown in the cumulative returns curve (%)
calmar	annualized_returns / max_drawdown
tail_ratio	Ratio between the 95th and (absolute) 5th percentile of the daily returns distribution. For example, a tail ratio of 0.25 means that losses are four times as bad as profits.

common_sense_ratio	tail_ratio(returns) * (1 + annual_return(returns))
total_pos_count	Total number of unique names held over the course of the sample period
max_pos_held	Maximum number of unique positions held in the sample period
mean_pos_held	Mean number of unique positions held in the sample period
median_pos_held	Median number of unique positions held in the sample period
pct_xnor_hedged	Percent of trading days the algorithm was in all cash or held one short and one long at the close
pct_days_invested	Percent of trading days the algorithm held an open position at the end of the trading day
median_hedge_ratio	median(short exposure / long exposure)
mean_hedge_ratio	median(short exposure / long exposure)
max_hedge_ratio	max(short exposure / long exposure)
min_hedge_ratio	min(short exposure / long exposure)
mean_max_long_pos_concentration	mean(daily largest long position % portfolio allocation)
median_max_long_pos_concentration	median(daily largest long position % allocation)
max_max_long_pos_concentration	max(daily largest long position % allocation)
mean_max_short_pos_concentration	mean(daily largest short position % allocation)
median_max_short_pos_concentration	median(daily largest short position % allocation)
max_max_short_pos_concentration	max(daily largest short position % allocation)

mean_median_long_pos_concentration	mean(daily median long position % allocation)
mean_median_short_pos_concentration	mean(daily median short position % allocation)
median_pct_std_of_allocations_by_name	median(standard deviation in % allocation for each name over sample period)
median_effective_pos_count	Median of a measure that estimates the daily number of effective positions. In a equally balanced portfolio will be equal the number of positions and reduces as the portfolio gets more skewed.
min_effective_pos_count	Minimum of the daily number of effective positions (see above).
mean_monthly_turnover	Average monthly turn-over in %
mean_profit_per_winning_trade	Mean % returns of all winning trades
median_profit_per_winning_trade	Median % returns of all winning trades
min_profit_per_winning_trade	Minimum % returns of all winning trades
max_profit_per_winning_trade	Maximum % returns of all winning trades
mean_loss_per_losing_trade	Mean % returns of all losing trades
median_loss_per_losing_trade	Median % returns of all losing trades
min_loss_per_losing_trade	Minimum % returns of all losing trades
max_loss_per_losing_trade	Maximum % returns of all losing trades
std_of_profit_per_name	Standard deviation in total profit generated by all traded names
decisions_per_day	Number of “round trips” (purchase/sale and subsequent sale/purchase of shares) divided by period trading days
user_backtest_days	Total number of days in all the backtests run on the

	algorithm prior to deployment

Works Cited

Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2014). Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance. *Notices of the AMS*, *61*(5), 458–471. doi:10.2139/ssrn.2308659

Bailey, D. H., Borwein, J. M., & Marcos, L. (2015). THE PROBABILITY OF BACKTEST OVERFITTING.

Bailey, D. H., & Lopez de Prado, M. (2014). The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality. *Journal of Portfolio Management*, *40*(5), 94–107. doi:10.3905/jpm.2014.40.5.094

Lo, Andrew W. "The Statistics of Sharpe Ratios." *Financial Analysts Journal* 58.4 (2002): 36-52.

P. Geurts, D. Ernst., and L. Wehenkel, "Extremely randomized trees", Machine Learning, 63(1), 3-42, 2006.

L. Breiman, "Random Forests", Machine Learning, 45(1), 5-32, 2001.

Sharpe, William F. "Mutual Fund Performance." *Journal of Business*, January 1966, pp. 119-138

Li, B., & Hoi, S. C. H. (2012). On-Line Portfolio Selection with Moving Average Reversion. Retrieved from http://arxiv.org/abs/1206.4626

Liu, Q., Lu, L., Sun, B., & Yan, H. (2015). A Model of Anomaly Discovery. *Ssrn*, *2015*(87), 1–79. doi:10.17016/FEDS.2015.087

Lin, M., & Chou, P. (2003). The pitfall of using Sharpe ratio. *Finance Letters*, (1997), 84–89. Retrieved from http://www.mgt.ncu.edu.tw/~chou/sharpe.pdf

Harvey, C. R., Liu, Y., & Zhu, H. (2016). ...and the Cross-Section of Expected Returns. *Review of Financial Studies*, *29*(1), 5–68. doi:10.1093/rfs/hhv059

Harvey, C. R., & Liu, Y. (2014). Lucky Factors. *SSRN Electronic Journal*. Retrieved from http://papers.ssrn.com/abstract=2528780\nhttp://papers.ssrn.com/sol3/papers.cfm?abstract_id=2528780

Lopez de Prado, M. (2013). The Probability of Back-Test Over-Fitting. *SSRN Electronic Journal*, *64*(s4 The Role of), S126–S130. doi:10.2139/ssrn.2308682

Kakushadze, Z., Lauprete, G., & Tulchinsky, I. (2015). 101 Formulaic Alphas. *Ssrn*, 1–22. Retrieved from http://ssrn.com/abstract=2701346

Schorfheide, F., & Wolpin, K. I. (2012). On the Use of Holdout Samples for Model Selection. *American Economic Review*, *102*(3), 477–481. doi:10.1257/aer.102.3.477

Beaudan, P. (2013). Telling the Good from the Bad and the Ugly: How to Evaluate Backtested Investment Strategies. *SSRN Electronic Journal*, (December 1999), 1–19. doi:10.2139/ssrn.2346600

Boudreau, K. J., Helfat, C. E., Lakhani, K. R., & Menietti, M. (2013). Performance Responses to Competition across Skill-levels in Rank Order Tournaments: Field Evidence and Implications for Tournament Design. *Working Papers -- Harvard Business School Division of Research*, *47*(1), 1–51.

Burns, P. (2006). Random portfolios for evaluating trading strategies. *Available at SSRN 881735*, (January), 1–16. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=881735

Christie, S. (2005). Is the Sharpe ratio useful in asset allocation?, (May), 1–48. Retrieved from http://www.mafc.mq.edu.au/linkservid/95558986-5056-AF00-5690347E9CC30E2C/showMeta/0/