

**Acknowledgements**

I want to thank my thesis advisor, Professor Yuxin Chen, for his patient and inspiring guidance over the whole course of this project. He provided me with many perspectives when my ideas were exhausted and pointed out my problems in a timely and accurate manner to stop loss in time. I want to thank Professor Marti Subrahmanyam, Christina Wang, Jens Leth Hougaard for their stewardship and dedication to the Business and Economics Honors program. I also want to thank our teaching assistant, Xinyi Yang, for organizing and facilitating every seminar in the entire academic year.

1. **Introduction**

With fewer regulation restrictions, higher leverage ceilings, the wider latitude of derivatives, and more flexible investment strategies, hedge funds can aggressively perform in both domestic and international markets to sprint illustrious active returns. Dramatically outperforming mutual funds since the revive from 1986, hedge funds have been viewed as the oasis of positive returns and become increasingly popular among accredited wealthy and institutional investors over the past decades. According to Barclays Hedge, the assets under management are estimated to have boosted from $950 billion to $ 12 trillion from 2000 to 2019, jumped by 1163%. As the growth in hedge funds has surpassed the overall growth of investment funds, the massive market has indicated unprecedented importance of the global hedge fund industry, arousing widespread interest in understanding its investment process as well as the various risk it faces.

A basic conflict between hedge funds and their investors is the extent to which they should publicly disclose their portfolio performance and composition. Hedge funds argued that the costs of frequent disclosure can outweigh the benefits. The revealing of priority information can cause free-riding by copycat investors or competitors on replicating trading strategies, depreciating their costly expenditures made to securities research or market analysis. For professional investors and speculators, more frequent disclosure will equip them with more comprehensive information to anticipate and thus trade ahead against the funds to capture the arbitrage, leading to higher purchasing prices for the funds (Wermers, 2001). The potential harms for hedge funds also include cutting down pre-tax returns for tax-management strategies and reducing the fund manager's incentive to explore and analyze.

As a result, hedge funds typically report performance on a monthly basis, and they sometimes may request confidential treatment to delay disclosure. Besides, as recapitulated by previous scholars, hedge funds often set barriers to capital inflows and outflows, such as subscription or redemption periods and lock-up periods, in the name of protecting managers from liquidity crisis (Li, Markov, and Wermers, 2012). The existence of such barriers confines investors' ability to respond to rapid market changes, such as the recent Coronavirus outbreak.

However, it is believed that high transparency is beneficial by allowing investors to make timely and informed decisions and reduce agency costs when managerial actions are opaque (Aragon et. Al, 2013). Have long understood the value of frequent disclosure of information, they also recognize that hedge funds may suffer from greater volatility when the market fluctuates. A hedge fund's portfolio can across multiple markets to contain multiple strategies, and such diversity has brought up complex risk correlations among macroeconomic variables and correlated flows that such strategies create. Therefore, timely disclosure can make investors better evaluate the magnitude of gains/losses when confronting big financial events, thus have a better sense on planning redemptions or adjusting capital allocations based on recent information, rather than waiting to the end of each month.

This controversy calls for the necessity to develop an approach to infer daily (high frequency) hedge fund returns when only monthly reported data is available. This paper introduces a factor-based replication method with machine learning models to compute daily hedge fund returns with little delay. The replication is based on publicly accessible daily return information of selected factors. Such daily monitoring, serving as a warning alarm, is especially important for investors during periods of high market volatility.

Previous related literature of hedge fund replication can be categorized into two main streams, the factor-based replication, and distribution-based replication. The former addresses more on path replication, aiming to capture risk factors driving hedge fund performance by running multi-linear regression from hedge fund returns on a series of style factors, see Gupta, Szado, and Spurgin (2008). The latter targeted to capture essential features of specific trading strategies, “referring to them as alternative beta strategies” (Glaffig, 2012) with little interest in duplicating the performance path anymore. The output of both streams is essentially a static portfolio, or an unchanged weight allocation, of the selected factors. By filling in the return of each asset class, they can get the proxy of the hedge fund return.

Even if it is a well-established area, few of the research has touched machine learning techniques in replicating returns. This paper’s model extends from the factor-based replication method, not only out of simplicity but also machine learning can remit some common pitfalls during the replication process. One of the issues of a factor-based model is over-fitting when too many factors are inputted. But by using Bagging and Regularization techniques in machine learning we can reduce overfitting when unstable estimation is generated. With the ability to quickly digest reams of data, machine learning can also play a role in factor selection from a large factor pool contributed by previous scholars.

What separates this model apart from the traditional ones also includes the goal. While most research aims to forecast long-term, i.e. monthly returns with past monthly data by creating a similar performing portfolio, we wish to construct a simple model to monitor *daily* hedge fund returns with daily, publicly available data.

To examine the robustness and feasibility of our model, we create two backtesting cases: case 1 covers from December 2nd, 2019 to April 3th, 2020, which includes a relatively peaceful first half and the last half with the market turmoil caused by COVID-19; case 2 covers from February 1st, 2008 to January 29th, 2009, which includes the fluctuating market conditions due to the 2008 financial crisis.

The remaining sections of the paper are organized as follows: Section 2 describes data and factor selection; Section 3 discusses the methodology; Section 4 exhibits the back-testing results and corresponding analysis; Section 5 concludes the reason for such model performances and the perspective for future improvements.

1. **Data**

**A. HFRX indexes**

Since we have no access to the real daily hedge fund return data, we use the indexes data, both monthly and daily, which is available from *Hedge Fund Research* database, to represent the real hedge fund data during the back-testing process to view the feasibility and robustness of our model. Considered the industry standard benchmarks of hedge fund performance, HFRX indexes utilize a rigorous quantitative selection process to represent the larger hedge fund universe of different trading strategies. we select six indexes, which cover all major hedge fund strategies and provide the most frequently reported performance on a day t+1 basis, with a common period back to January 2004.

Another reason for using indexes is that learning and replicating the history does not necessarily provide insights in true nature of the underlying assets' performance, as there are many idiosyncrasies, such as the dynamics involving carry aspects and tail event behavior (Glaffig, 2012) or specific trader’s talent can dominate the performance of a certain strategy during a certain period. But for hedge fund indexes, idiosyncrasies can be diversified away to some extent, and the remaining systematic characteristics would dominate and thus have a more meaningful and feasible out-of-sample replication.

**B. Factor Data**

Identifying a set of factors as comprehensive but still least independent as possible to capture hedge fund return characteristics under different market conditions can be very challenging. Previous scholars have contributed to a rich pool of market indexes and factors that covers different major markets, regions, sectors, and currencies. Among all, Fung and Hsieh (2004) go beyond traditional asset indexes and have found eight common asset-based style factors[[1]](#footnote-1) that can explain up to 80% of monthly return variations. In this paper, out of the twenty candidate factors in the pool, each hedge fund index is run against all possible combinations of the candidate factors, and the combination with the highest explanatory power is considered as the most significant factor set to be used for this strategy in this paper. Overall, thirteen factors in total are left after the separate six selection process. The data of each factor is downloaded from *Yahoo Finance*, *MSCI*, and *CBOE* website.

Below is the table of the factors for replication:

|  |  |  |
| --- | --- | --- |
| Market | Category | Factor indexes |
| Equities | Broad Market | S&P 500; Willshire 5000 |
|  | Size spread | (Russell 2000 - S&P 500) |
| Fixed Income | Broad Market | Treasury 10 Years;  ICE BofA 1-3 U.S. Corporate |
|  | Size spread | (Moody‘s Baa Yield - Treasury 10 Years) |
| Emerging Market |  | MSCI Emerging Market ETF |
| Trend Following | Bond | Return of PTFS Bond Lookback Straddle |
|  | Currency | Return of PTFS Currency Lookback Straddle; USD/JPY |
|  | Commodity | Return of PTFS Commodity Lookback Straddle |
| Volatility |  | CBOE VIX |
| Options |  | CBOE BuyWrite (BXM) |

1. **Methodology**

The proposed daily prediction model involves two steps: first, train the model to let it learn the projection from monthly factors onto the monthly hedge fund returns,

where is the hedge fund monthly return at time t, is the monthly return of the i-th factor; Second, we input the daily information of each factor into the trained model and the output is the projected daily hedge fund return.

Machine learning models demonstrate outstanding advantages in inferring patterns and fitting projections from history on future data. In this paper, I choose Random Forest and Support Vector Machine Regression as they have a totally different mathematical mechanism. Random Forest is an ensemble machine learning technique capable of performing both regression and classification tasks using multiple decision trees and a statistical technique called bagging. Each tree utilizes a subset of data and factors and breaks down the selected dataset into smaller and smaller subsets until it forms a leaf node with a set of “if-then-else” decision rules. The final numeric result is the mean of all involved decision trees. Support Vector Machine Regression (SVR) attempts to construct a hyperplane and decide a decision boundary at a certain distance from the hyperplane such that all data closest to the hyperplane or the support vectors are within that boundary line, thus fitting the line with least error.

1. **Interpretation of Empirical Result**

The training and validation set starts from January 2004 to November 2019. Below is the summary table of the training and validation mean square error (MSE), which is the loss function used to evaluate the prediction power of our models. Among all, Random Forest outperforms SVR in every case. Also, except for CTA, all other five strategies have their validation MSE better than the baseline MSE, demonstrating a proper selection of the factor set.

**Train/ Validation Mean Square Error Summary**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Trading  Strategy | Random  Forest  (Train) | Random  Forest  (Val) | Support Vector  Regression (Train) | Support Vector  Regression (Val) | Baseline |
| Convertible  Arbitrage | 2.0902 | 0.3396 | 1.0452 | 0.4151 | 0.4607 |
| Event Driven | 0.7316 | 1.0154 | 1.1941 | 1.0495 | 2.1389 |
| Equity Hedge | 0.8219 | 0.8365 | 1.4733 | 1.0978 | 1.3547 |
| Equity Market  Neutral | 0.8967 | 0.5201 | 1.0462 | 0.6129 | 0.7245 |
| CTA | 3.4172 | 2.9550 | 2.4627 | 3.3586 | 2.3132 |
| Merger  Arbitrage | 0.3403 | 0.6130 | 0.3249 | 0.6521 | 0.7119 |

To test the feasibility and robustness of our model, we set two back testing cases: case 1 starts from December 2nd, 2019 to April 3th, 2020, which includes a relatively peaceful first half and the last half with the market turmoil caused by COVID-19; case 2 starts from February 1st, 2008 to January 29th, 2009, which includes the fluctuating market conditions due to the 2008 financial crisis. Therefore, case 1 is an out-of-sample test locating out of the training process, and it’s more appropriate to see if the model can indeed be used for future prediction. Case 2 is an in-sample test, which locates within the training process so that it can examine if this model can precisely analyze what your portfolio has gone through every day in the past.

Below are the summary tables of the testing MSE and their corresponding baseline MSE in both test cases.

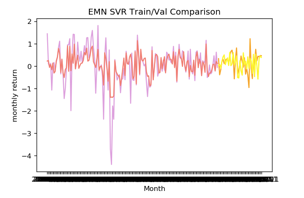
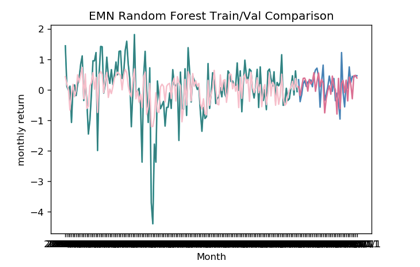
**Mean Square Error Summary of Case 1**

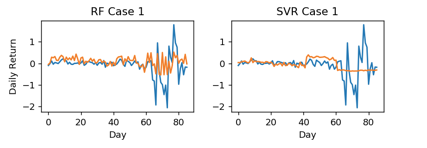
|  |  |  |  |
| --- | --- | --- | --- |
| Trading Strategy | Random Forest | Support Vector  Regression | Baseline |
| Convertible Arbitrage | 0.5195 | 0.3363 | 0.2268 |
| Event Driven | 0.6289 | 0.4232 | 0.3445 |
| Equity Hedge | 0.8914 | 0.5249 | 0.4923 |
| Equity Market Neutral | 0.3675 | 0.2866 | 0.2288 |
| CTA | 0.2405 | 0.9857 | 0.2159 |
| Merger Arbitrage | 1.4342 | 2.0655 | 1.8267 |

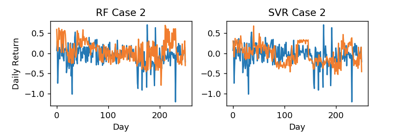
**Mean Square Error Summary of Case 2**

|  |  |  |  |
| --- | --- | --- | --- |
| Trading Strategy | Random Forest | Support Vector  Regression | Baseline |
| Convertible Arbitrage | 1.0976 | 1.2551 | 1.1449 |
| Event Driven | 0.6873 | 0.6159 | 0.3281 |
| Equity Hedge | 1.4489 | 0.9121 | 0.4620 |
| Equity Market Neutral | 0.2339 | 0.1823 | 0.1500 |
| CTA | 0.5878 | 1.0581 | 0.3214 |
| Merger Arbitrage | 0.3497 | 0.4805 | 0.4242 |

1. **Equity Market Neutral**

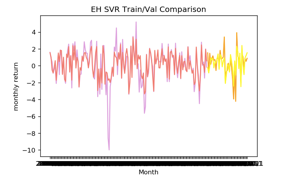
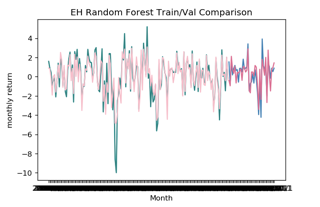


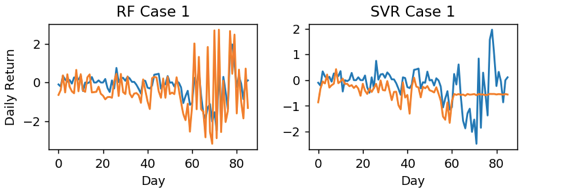


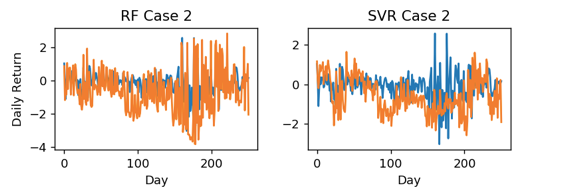


Above are the plots of training, validation as well as testing results in both cases of the Equity Market Neutral strategy, which is designed to hold equal market value in both long and short positions to have a net-zero risk exposure. Both validation MSE is better than the baseline. Random Forest and SVR both perform fairly in the first half in case 1, which is a relatively peaceful period. But after the COVID-19 roiled global equity markets, Random Forest, although successfully captures the volatile trend in general, has lost the accuracy of mimicking the magnitude of each spike. SVR basically loses its predictive power as it generates a stationary output far different from the true, highly volatile trend. As for case 2, Random Forest outputs a decent replication in the middle part, which is the relatively stable period compared to the fluctuating market conditions after August 2008; while SVR oversteps the entire period.

1. **Equity Hedge**



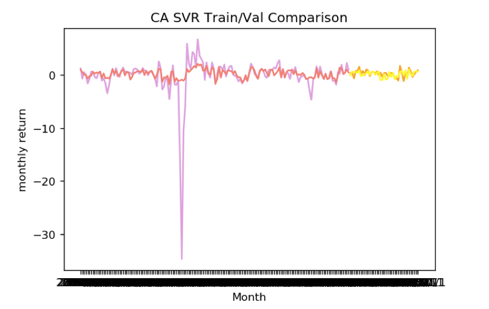


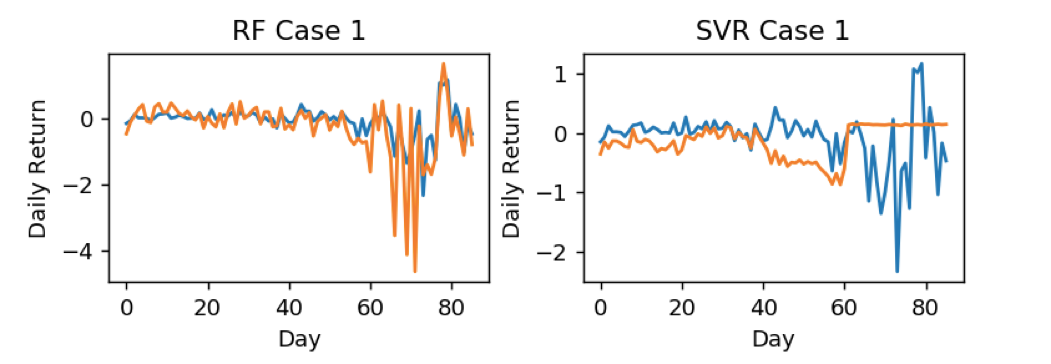


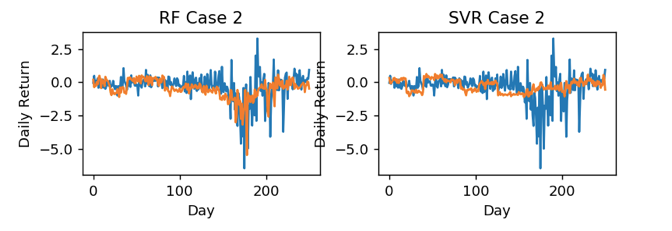
Above are the plots of training, validation as well as testing results in both cases of the Equity Hedge strategy, which invests in both long/short positions in competing assets to win double profit. Both validation MSE is better than the baseline. In test case 1, the prediction given by Random Forest of certain peaks is way greater than the actual data in terms of the magnitude, especially after the US stock market experienced several circuit breakdowns after March 2020. But it is worth noting that Random Forest gives a trend forecasting with high accuracy within these highly volatile months. In contrast, SVR gives a prediction that is close to a straight line in the turbulence of the coronavirus outbreak. In case 2, the trend of Random Forest prediction is closer to the real situation, but the amplitude is much higher, while SVR is much smaller.

In fact, hedge funds with the greatest exposure to equity markets have encountered the biggest opportunities and challenges in the past months. They can benefit from using short positions as they did in the past crises. However, according to the HFR database, hedge fund investors have withdrawn $33bn in Q1 2020, the highest quarterly redemption number since the 2008 crisis. Such redemption can cause short squeeze and liquidity crisis for fund managers, thus making it harder to meet the margin call and imposing a longer time to rebalance their net positions. Therefore, the performance of equity hedge funds and equity market neutral funds are full of manifold uncertainties when faced with the financial crisis, thus very difficult to imitate and predict their trends and magnitudes.

1. **Convertible Arbitrage**

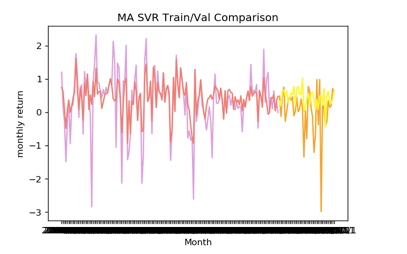
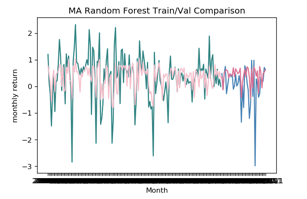


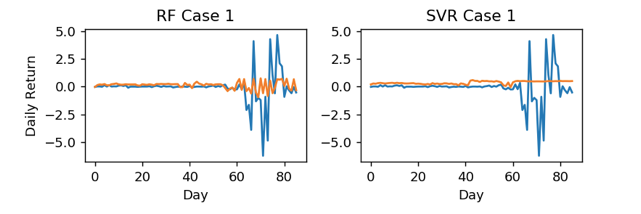


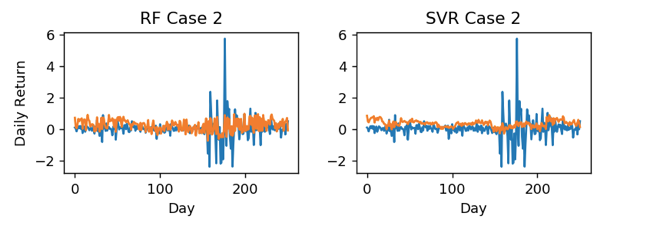


Above are the plots of training, validation as well as testing results in both cases of the Convertible Arbitrage strategy, which attempts to exploit price anomalies when longing convertible securities and shorting the same issuer’s common stock. Both validation MSE is better than the baseline. In case 1, Random Forest does a great job overall except for it over-estimates the loss during the market downside caused by COVID-19, while the prediction results of SVR are far away from the true value. In case 2, Random Forest has very desirable replication of the daily convertible arbitrage hedge fund return, especially accurate in forecasting several important market declines during the 2008 financial crisis, while somewhat under-estimates the high peaks. The overall prediction power of SVR is still very unsatisfactory.

1. **Merger Arbitrage**

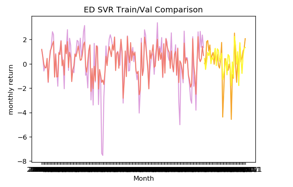
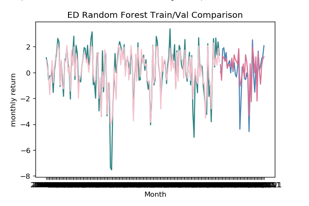


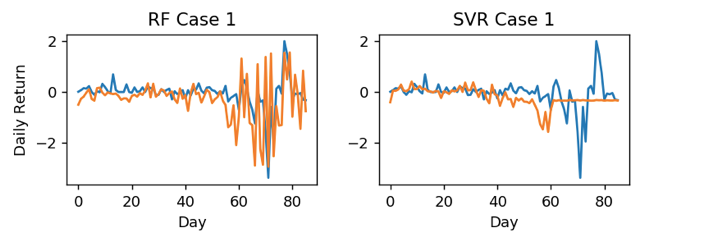


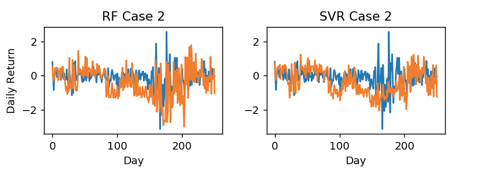


Above are the plots of training, validation as well as testing results in both cases of the Merger Arbitrage strategy, which simultaneously go long the “target” company and sells short the “acquiring” company to create "riskless" profits in a merger and acquisition deal. Both validation MSE is better than the baseline. In test case 1, the peaceful period from the end of 2019 to January 2020 has few deals, but the advent of the coronavirus has broken many capital chains and pushed many companies to the verge of bankruptcy, leading to a surge in M​​&A deals. In both Case 1 and Case 2, although the performance of the SVR is not very satisfactory, the up-and-down trend forecasted by Random Forest is almost consistent with the real data, but the numeric value is way too conservative than the reality. It is worth mentioning that the performance of merger arbitrage funds depends on whether the deal passes and the price when the deal passes, therefore closely related to the regulation and policy of the industry. This policy information is difficult to be reflected in the factors that we find to represent the market, so it inevitably has a certain impact on the accuracy of prediction.

1. **Event Driven**

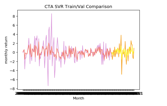
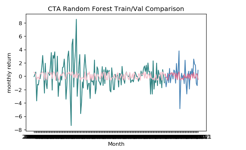


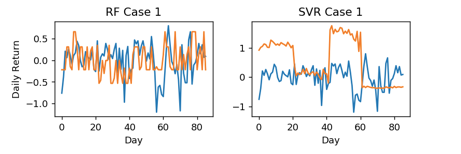


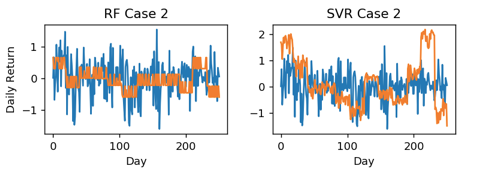


Above are the plots of training, validation as well as testing results in both cases of the Event Driven strategy, which seeks to exploit pricing inefficiencies that may occur before or after an event, specifically referring to distress securities investing here. Event driven strategy consists of buying distressed securities from a company subject to bankruptcy or restructuring at a discounted price and sell them when they appreciate again after the restructuring process. Both validation MSE is better than the baseline. Random Forest performs fairly in case 1, except for the bias regarding the magnitude of each spike. But in the other three plots, we see that neither Random Forest nor SVR performs ideally in those cases. As the restructuring and liquidation process may take months or years to complete, this strategy is very sensitive to changing market conditions. Hence, distressed security investing has a significant correlation with major bond and stock indexes. Very similar to the aforementioned equity hedge and equity market neutral strategy, it contains too many uncertain factors, piling up a lot of pressure on the model to learn patterns and make predictions.

1. **Commodity Trading Advisor (CTA)**







Above are the plots of training, validation as well as testing results in both cases of the CTA strategy, which generally uses futures contracts to achieve the investment objective, covering global listed financial, commodity and foreign exchange markets. This is the only strategy whose both validation MSE is worse than the baseline, indicating the misspecification of factors. As a defensive strategy that is typically uncorrelated to public markets, even the most related factors, such as commodity and currency trend-following factors seem to deliver little information here. Therefore, the prediction results in both test cases can exhibit strange patterns.

1. **Conclusions**

This paper introduces a factor-based replication approach with machine learning models to infer daily hedge fund returns using daily information of selected factors. On the whole, from the perspective of model picking, the prediction results of Random Forest are far superior to SVR in both in-sample and out-of-sample tests. The accuracy of Random Forest’s prediction on the overall fluctuation trend is very high, and the accuracy of the magnitude prediction is relatively lower. By contrast, SVR has little value in predicting either the trend or numerical return, as it typically generates stationary output even if the real market is extremely volatile. The main reason for such discretion between model performance can be explained by the model mechanism. Random Forest is an ensemble model that consists of several decision trees, each trained on a subset of observations and factors such that every tree is made least correlated, and the model regression output is the mean of each tree. As the factors used in this paper can be somewhat inter-correlated regarding the complex relationship among capital markets, the random forest performs better in terms of eliminating intercorrelation and autocorrelation. Moreover, the bagging technique employed in Random Forest can automatically reduce the variance of predictions. The performance of SVR may be improved by incorporating some categorical features as input, such as dummies representing regions or sectors.

From the perspective of the replication accuracy score, except for CTA, all the other five strategies have proven such factor combinations are reasonable and effective. But in most cases, although the predicted trends are correct in general, the predicted value of return always differs from the real data, especially when the market is experiencing turmoil. There are several possible reasons: First, apply the mapping relationship between the low-frequency data onto the high-frequency data does produce a large error, because the low-frequency data itself omits or hides a lot of information. Second, given the late inception date of the HFRX index database, the six strategies that can be found only record data from January 2004, so the limited amount of data can create large bias. The third reason is the wrong assumptions of the linear model. Some scholars have detected the nonlinear relationship between hedge fund returns and the returns of the traditional assets underlying their trades, typically given by the fact that hedge fund portfolio managers change their exposures to the factors according to their view or expectations towards the global economic conditions (Jaeger, 2008). These kinds of idiosyncrasies, such as manager's trading style, mindset, and talents, as well as synergy effect generated in multi-strategy funds, sometimes can dominate the other factors within the performance characteristic and make the whole process no more static and stationary, thus very intractable to model it in detail.

Therefore, to improve the performance of the model, we need to be more rigorous in choosing the factors. We can also create some categorial features such as dummy variables to improve the accuracy of SVR and more models. At the same time, considering the dynamics of the hedge fund returns, we can create a lag on the VIX index, or any other good estimates of investors’ perception towards the overall market, to incorporate manager’s “tactical trading behavior”. Furthermore, we can run the factor selection procedure every month to incorporate the dynamics in hedge fund risk exposure over time.

Even if the current models cannot deliver impressive performance due to the aforementioned reasons, the idea and framework itself still have potential applications. By more elaborately identifying factors, it can serve as the early warning system for investors to make in-time redemption decisions; Investors can also hedge against the unwanted risk by taking corresponding positions if it helps detect abnormal return signals. Besides, this generic method can be extended to more asset classes with low disclosure frequency.

**Work Cited**

Aragon, George O., et al. “Why Do Hedge Funds Avoid Disclosure? Evidence from Confidential 13F Filings.” *Journal of Financial and Quantitative Analysis*, vol. 48, no. 5, 2013, pp. 1499–1518., doi:10.1017/s0022109013000483.

Wermers, Russ. “The Potential Effects of More Frequent Portfolio Disclosure on Mutual Fund Performance,” *Perspective*, 2001

Li, Daniel, et al. “Monitoring Daily Hedge Fund Performance When Only Monthly Data Is Available.” *SSRN Electronic Journal*, 2009, doi:10.2139/ssrn.1362265.

Gupta, Raj, et al. “Performance Characteristics of Hedge Fund Replication Programs.” *SSRN Electronic Journal*, 2008, doi:10.2139/ssrn.1524293.

Glaffig, Clemens H. “Insight—Distributional Hedge Fund Replication via State Contingent Stochastic Dominance.” *Hedge Fund Replication*, 2012, pp. 76–89., doi:10.1057/9780230358317\_6.

Fung, William, and David A. Hsieh. “Hedge Fund Benchmarks: A Risk-Based Approach.” *Financial Analysts Journal*, vol. 60, no. 5, 2004, pp. 65–80., doi:10.2469/faj.v60.n5.2657.

Jaeger, Lars. “Alternative Beta Strategies.” *Partners Group Research*, 2008.

1. Data available at http://faculty.fuqua.duke.edu/~dah7/DataLibrary/TF-FAC.xls [↑](#footnote-ref-1)