Postgraduate Diploma in Financial Analytics and Algo Trading (Module 6: Algo Trading and Quantitative Investment Strategies)

Student name: Chan Ching Kit
Student number: 20242725

All-round Analysis in Automation

Ching Kit Chan

Abstract

This paper explains the project of mimicking a hedge fund to apply algorithmic trading strategies. The project aims to discover profitable algotrading strategies that are aligned with the prescribed fund's mandate, in both top-down and bottom-up approach. To achieve a more realistic result, details in the operation such as the architecture of the firm and regulations will be discussed, and the programming codes created are targeting for real trading.

Introduction

In the current digital age, the competition between trading firms is increasing exponentially as new technologies emerge. One of the breakthroughs of technology is the machine learning model, where the emergence of transformer drastically improves AI's capabilities. And that has opened huge possibilities for different industries to take advantage of AI and improve their business, such as improving their efficiency or lowering their cost.

While many companies are starting to experiment with AI or even implementing it, the financial industry has been implementing it for a long time. However, AI is still only a tool and cannot replace the work for analysts.

The objective of this project is to discover the possibility of replacing the analysts and the operation of a hedge fund with coding and machine learning. Several techniques such as big data handling, web scraping and machine learning are applied to overcome the challenge for implementing the idea. All the data processed in the project are clean in such a way that no lookahead bias occurred and the details for live trading are considered, therefore, the reliability of the result would be high. Moreover, the regulatory is researched to mimic the procedure of a hedge fund.

Literature review

In the book *Advances in Financial Machine Learning* (Prado, 2018). He mentioned a time decay method for machine learning model training. The time decay refers to the process of assigning diminishing weights to older observations compared to newer ones in adaptive market systems. The decay function is determined by a user-defined parameter, "c," which determines the rate of decay.

The function provided in the book considers the cumulative uniqueness of observations and assigns weights accordingly, with the newest observation receiving a weight of 1 and the oldest observation receiving the lowest weight. The cases to consider are:

c = 1, which means there is no time decay, and all observations have equal weight.

0 < c < 1, which indicates linear decay over time, but every observation still has a positive weight.

c = 0, where weights converge linearly to zero as they become older.

c < 0, where the oldest portion of observations (cT) receives zero weight and is effectively ignored.

This time decay function is referenced in the machine learning training in this research since it makes sense that the more recent data should have more value for the model to learning. Therefore, calculating this time decay allows for better way to sample the data for training. However, the approach to calculating time decay is slightly different than Prado since the dataset for the model is different. Prado also considered the uniqueness of each row since he is handling data that are highly overlapped. In this project, the row of the data does not overlap with others, so the time decay is simply linear, without considering the uniqueness.

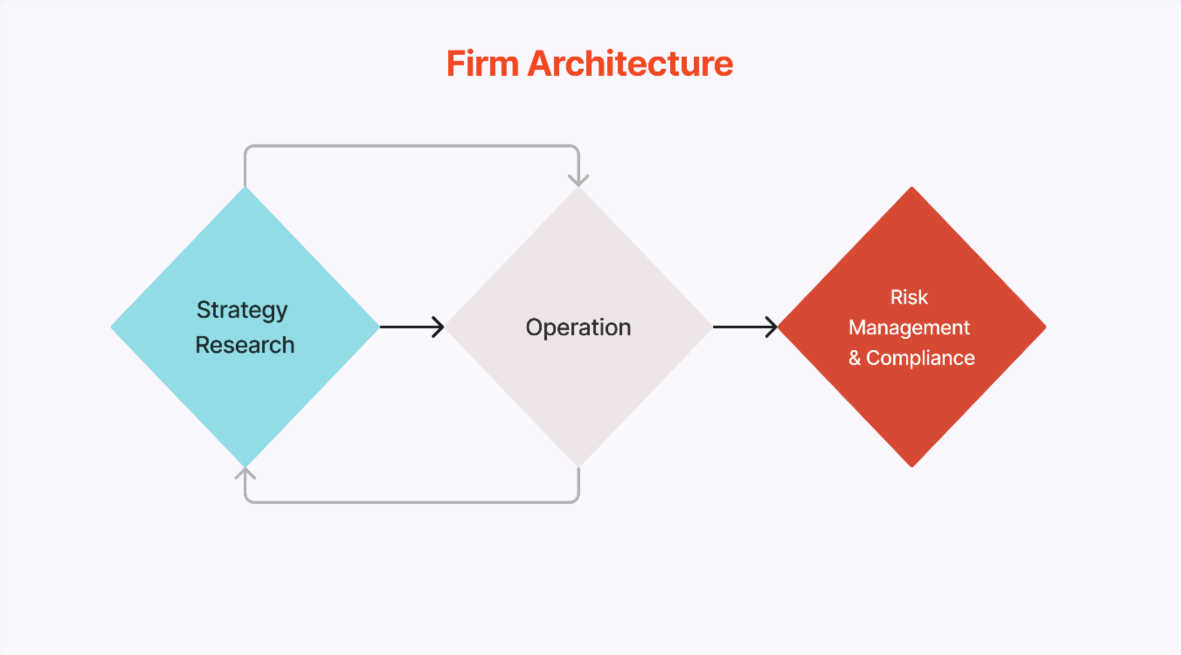
Research Methodology

As mentioned, the research is conducted with the notion of mimicking hedge funds. Therefore, the research methodology of the project is structured in a way that considers different departments in the firm so that a realistic result will be generated.

- Starting from the firm’s architecture and objective

Although the regulation on hedge funds is looser than other financial institutions such as pension funds, the freedom it provides is also making hedge funds prone to higher risk.

Therefore, the architecture and the strategies of the firm must be well-designed to avoid any systematic risk.



As the figure above shown, there are three main segments in the architecture. While some big hedge funds might have more segments and more complicated architecture, these three segments are the basic and it is enough to operate the firm to develop strategies to trade while amending the regulatory. In an architecture like this, the segments of strategy research and operation act as the brain of the firm to seek revenue, and the risk management and compliance segment will oversee finalizing the decision.

1) Strategy Research

It is important for the fund to define its mandate before conducting any research since different styles of strategies suit different return target and risk profile. For example, risk-averse funds would prefer a lower risk strategy such as arbitrage. However, strategies like this are not likely to generate remarkable profit. Some fund would consider risk like short squeeze which we saw on GME in 2021 and decide to not allow for short position. And some funds would consider the liquidity problem and decide to not trade several markets such as emerging markets.

For hedge funds, it has a high degree of freedom in its trading strategies and the return target and risk profile can vary a lot between funds. The risks to be considered are up to the decision of the firm. While the return is not the focus in the project, some general wisdom is amended for a better chance to get a better profit. Wisdom such as not to trade illiquid market which is likely to have huge slippage when the firm executing order, manage the leverage and exposures, etc.

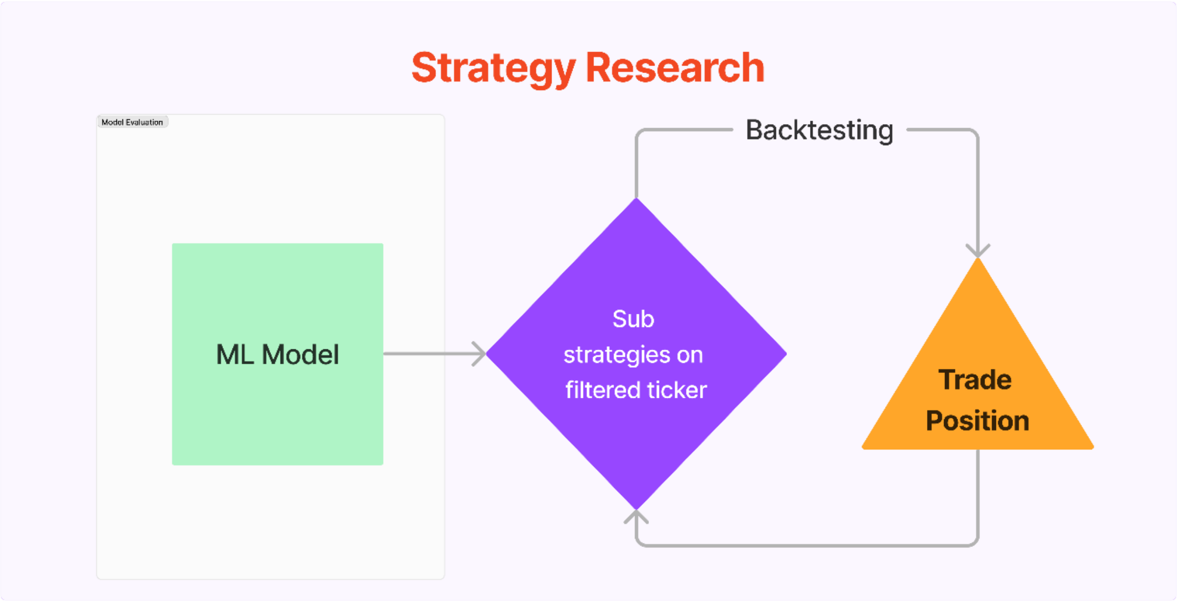
Therefore, the objectives for the hedge fund in this project is defined as this:

- Target for absolute return instead of beating the market.
- Only trade the stocks within the S&P500 which are very liquid.
- Short positions are allowed.

And the direction of the research will follow these objectives.

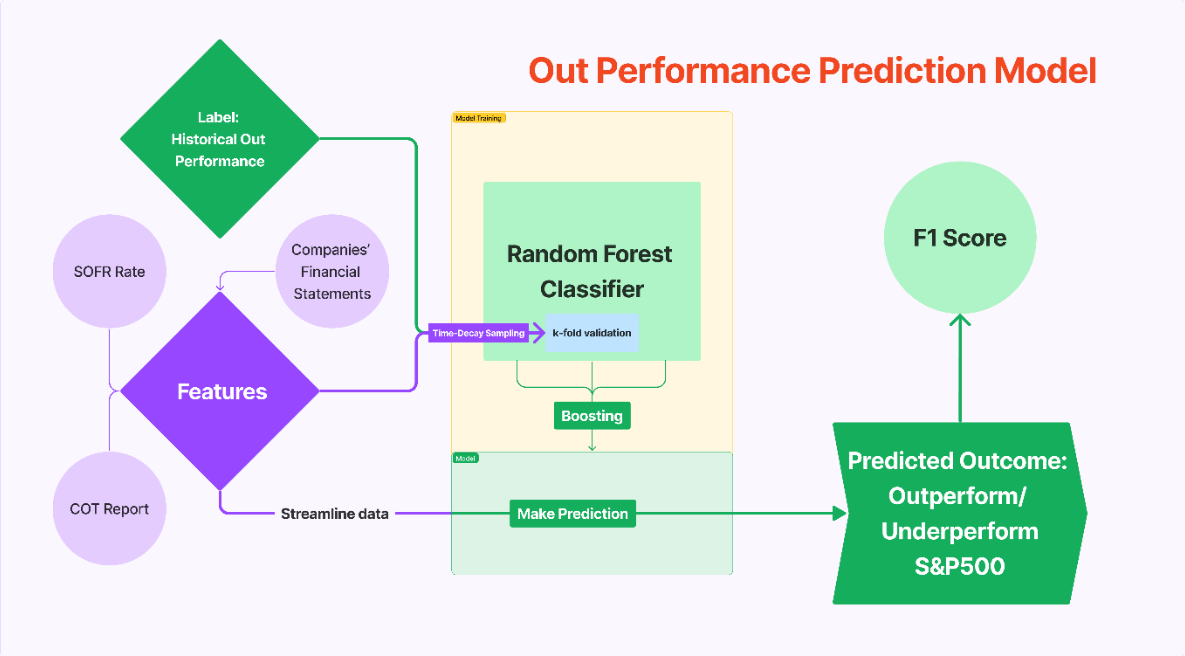
1.1) Strategy structure

A well-established strategy structure allows for faster strategy development since the researchers in the firm can separate to focus on the part that matches their strength.



As the figure above shown, the strategy consists of a machine learning model and sub strategies, and a backtesting process for evaluate the strategies created. The machine learning model will filter the stocks for the sub strategies to be applied on. In the best scenario, the machine learning model and sub strategies would benefit from each other, and the edge of the strategy would be compounded and create a bigger edge for the strategy. Also, this structure eliminates the risk of leaking strategies, this will be explained in the risk management section.

1.2) The machine learning model



The flow chart illustrated the architecture of the machine learning model. It is a random forest classifier with adaptive boosting which aims to filter out the stocks that will be outperforming the S&P500. Instead of working just like some rules for stocks filtering, the goal for the model in the strategy is to act as an analyst to analyze the market. Usually, stock analysts will analyze the stock with both macro and micro analysis, where macro accounts for the entire market and the micro account for the financial performance of the company. Since it is unlikely that stocks price can continue to gain under a bad business environment, and a company with unsatisfying financial is unlikely to outperform other stocks that are better, this all-round analysis is the preferred way for stock analyst.

1.2.1) Dataset

The features chosen for the machine learning model are a combination of macro and micro data in the period from 1/10/2020 to 1/1/2024, approximately 3 years and 1 quarter of data. The macro data consisted of the Secured Overnight Financing Rate (SOFR) & the ratio derived from the Commitment of Trader Report (COT) and the micro data consisted of the mathematically adjusted ratio derived from the financial statements of the S&P500 companies. The COT is published by the CFTC, it discloses the long, short, spread positions held by different categories of traders in the futures market. The features derived from COT are the data from E-mini S&P500 futures which gives insight on the US stock market. The details of all the features are as follow:

- The SOFR: Calculated as a volume-weighted median of transaction-level tri-party repo data collected from the Bank of New York Mellon as well as GCF Repo transaction data and data on bilateral Treasury repo transactions cleared through FICC's DVP service, which are obtained from the U.S. Department of the Treasury’s Office of Financial Research (OFR)[1]. It is considered forward-looking because it is published daily and represents the rate at which overnight borrowing is expected to occur in the future. It provides market participants with an indication of the prevailing borrowing costs for secured overnight funding. And the combination of median makes it a more resilience benchmark rate than others.
- Buy Side Dominance (COT): The Asset Manager/ Institutional and Leveraged Funds are the so-called buy side in the market which are more likely to make directional bets on the price. This feature, calculated by summing the long and short positions held by the buy side and then divided by the total open interest, may related to different characteristics in the market since there are more informed traders instead of retail traders.
- Asset Manager/ Institutional Bias (COT): The bias can be calculated by dividing long positions with the sum of long and short positions, the number is a value between 0 to 1 where a higher value represents more long-bias, and vice versa.
- Leveraged Funds Bias (COT): Same as above with different entity.
- Nonreportable Bias (COT): Same as above with different entity.
- Adjusted P/E (Income Statement): The P/E ratio, a measure of valuation referencing the earning, is calculated by dividing the price by the EPS value. A low value means the company might be undervalued. Since the calculation of it tends to produce extreme value due to the division of small numbers, an adjustment is made to eliminate the impact of the extreme value to the machine learning model. The logic is: If the value is not within the range of -1 to 1, log the value with base 10, otherwise keep the value as the same.
- Adjusted Quick Ratio (Balance Sheet): The quick ratio is a measure of the health of the company’s balance sheet for the upcoming few months. It is a short-term measurement that matches the model as the prediction for the model is the next 3 months which is also short-term. Usually, a low quick ratio represents a higher risk for bankruptcies, which might trigger a sell-off. The measurement is also adjusted with the same transformation as the

[1] https://www.newyorkfed.org/markets/reference-rates/sofr

Adjusted P/E.

- Free Cashflow per share (Cashflow Statement): The free cashflow per share is the free cashflow divided by the amount of the shares so that it is comparable between different stocks. It represents the available funds for the companies to utilize after all the expenses. A high value shows that the company has excessive funds to invest in new projects and it is generally beneficial to the stock price.

With the features selection as this, the model is expected to learn the conditions of the companies and the views from the market participants toward the US market. This information is important for predicting whether a stock can outperform the market.

The label of the dataset, the out performance, is calculated by subtracting the market return from the stock return in the next 3 months. Since the model is a classification model, the value of the out performance must be transformed into the sign of the number.

1.2.2) Training and validating

Before the model training, a time decay weighting and the dataset must be preprocessed. The time decay weighting is calculated with the function written by Marcos Lopez de Prado which is explained in detail in the literature review section. Then features in the dataset are scaled with normalization where every datapoint in the features is scaled with the maximum and minimum value in the column. The dataset is sorted according to the time and sliced into training, validation, and test sets. Shuffle is avoided in the splitting of the dataset to avoid any lookahead bias where the data supposed to be in the future data is learned by the model for predicting the data in the past.

The training of the model tokes the 80% of the earliest data and validate with the F1 score which is a balanced metric that considered both precision and recall. It was processed with a grid search hyperparameter optimization which iterates and trained multiple models with all the possible combinations in a set of hyperparameters. And for each model, it is trained with the k-fold cross-validation which averages the performance over multiple iterations, providing a more reliable estimate. After all the training and optimization, the best estimator is created with the F1 score of 0.502.

1.3) Sub Strategies

The sub strategies are the next step for the trading algorithm to make decisions.

However, it does not have to be researched only after the model is built. In fact, it can be researched concurrently with the model once the predicting label of the model is decided. Two sub strategies are built in this project, which are the Smart Beta and Breakout strategies. Referencing from popular stockbroker, the commission of each trade is set to 0.35 USD and the initial capital is set to 100 (which directly visualizes the percentage of the capital), with each order betting 50% of the initial capital.

1.3.1) Smart Beta

This strategy is like a smart beta ETF which has some criteria to select the better performing stocks. The logic for this strategy is to buy and hold 3 months for the stocks that the model predicted to outperform S&P500. This strategy is directly benefited from the prediction of the model since the model predicts the difference between the stock return and S&P500 return. If the accuracy of the model is high, running the strategy will build a portfolio with stocks that can beat the market. Therefore, this model relied heavily on the performance of the machine learning model. The strategy will be underperforming the S&P500 or even result in a loss if the prediction from the model is inaccurate. Still, the simplicity of the strategy is making it easy to build and maintain. It can be easily combined with other strategies without making the logic too complicated and not interpretable. Moreover, the gross exposure is limited since the strategy will only have 1 long position at most for each stock in the S&P500. So, the tail risk associated with an extreme situation can be estimated.

1.3.2) Breakout

While the smart beta directly benefited from the prediction of the model, the breakout strategy is not since shorting position is allowed in this strategy. But this strategy is still being researched under the assumption that the stocks outperform the market might have some specific characteristics that the breakout strategy can benefit from. The trading logic is to trade when the price breakout from consolidation. A long signal will only be generated when the price is above 21 periods ema, and it is a short signal when the price is below the ema. And taking profit or stoploss based on the weekly volatility and risk-to-reward ratio, which is 1:1 for the simplicity for research.

This strategy incorporates technical analysis concepts such as key levels, support and resistance, and breakouts. In behavioral finance, individuals often make different decisions based on different trends and when approaching key levels, such as breaking out upward or downward. It aims to benefit from strong price movement. The notion of it is simple: when near support or resistance, there is assumed to be price rejection which is hard for the price to breakout that key level. Therefore, the strategy speculates on the price continuing the same direction after price breakout the key level since it demonstrates strong buying/selling interests. It can be explained in behavioral finance as the market participants are making decision when they observed abnormal price movement, a strong price action might strengthen their bias and move the price even further to the direction of breakout. So, under this assumption, it is a profitable opportunity if the algo can buy/sell as soon as a breakout is confirmed.

Unlike the first strategy, this strategy is much more complex. Although it does have the downside of being harder to build and maintain, it has more parameters which allow for parameter optimization which can potentially improve the performance. Since it is using the weekly volatility for take profit and stoploss, it has a shorter period of holding the stocks which should give more volatility to the strategy. While it makes sense to set up a position limit to avoid gross exposure inflated unlimitedly or overexposed to a specific stock, it prevents the later generated signals from having an impact on the portfolio. But in fact, the later signals should be as important as the earlier signals. Therefore, this kind of bias toward earlier signals must be avoided to evaluate the strategy fairly. Not limiting the position is also allowing for observing the frequency of the signals occurring. Lastly, the strategy is allowing for shorting stocks so there is also a risk of a short squeeze like the GME in 2021 which the loss cannot be limited by the position size.

1.3.3) Combined strategy

This is not a separate strategy but an idea to combine different strategies based on their characteristics and thus achieving a better result. The logic is to allocate funds to the existing strategies and optimize for an allocation that performs the best. Combining strategies allow the strategies to hedge each other, achieving a smoother equity curve and more diversified portfolio. And it is also more diversified in terms of trading logic, which is beneficial since the market is not likely to favor one style of strategy forever and there must be a period of drawdown even for the best strategy. Currently the research on this combined strategy is limited due to the lack of available strategies. The combined strategy researched in the project is simply allocating 50% for both developed strategies.

1.4) Performance metrics

There are several performance metrics, all calculated in rolling 3 months and full period, for analyzing the strategies return thoroughly. Including CAGR/ Mean drawdown, Sharpe ratio, Beta and Alpha.

1.4.1) CAGR/ Mean drawdown is a straightforward measure of profitability with the risk accounted. Instead of using the more popular maximum drawdown as the denominator, using mean drawdown can avoid the problem of maximum drawdown being a value that will only increase throughout the periods instead of adapting.

1.4.2) Sharpe ratio, named after William F. Sharpe, measures the risk-adjusted return of an investment or portfolio with the consideration of risk-free rate and volatility. A positive Sharpe ratio shows that the portfolio has a higher return than the risk-free rate, and a higher value indicates a better risk-adjusted return.

1.4.3) Beta is a measure of an investment's sensitivity to the overall market movements. A beta of 1 indicates that the investment tends to move in line with the market. A beta greater than 1 suggests that the investment is more volatile than the market, while a beta less than 1 implies lower volatility. A negative beta means the investment moves in the opposite direction of the market.

1.4.4) Alpha measures the excess return of an investment or portfolio compared to its expected return based on its beta and the market return. It is used to evaluate the performance of the strategy compared with its estimated return.

2) Operation

The operation segment of the firm will be doing most of the programming work, cooperating closely with the strategy researchers, retrieving all the data needed for research purposes upon researchers' request. The real-time data streamlines will be built for any strategy that delivered promising backtest and forward test results. Python and its packages such as Pandas and NumPy are utilized for computational work. Moreover, this segment oversees building and maintaining a system that handles trade signals for both paper trade and live trading, monitoring the portfolio and logging all the trades. There are 4 parts for the operation segment to work on in this project.

2.1) Handling the data

To allow for the research to be conducted, it is essential to retrieve, structure and transform the raw data. Firstly, the API from Financial Modelling Prep is connected to retrieve the historical S&P500 constituents. Then retrieve the financial reports, price, and market cap of the constituents right before the start date of the research period and calculate all the micro financial ratios needed. Mathematic transformations are done on P/E and quick ratio where all the value over -1 to 1 will be transformed by logarithmic with base 10, the absolute value of the negative value will be transformed and multiply with -1 since the logarithmic cannot be applied on negative value. All of these are being done in the final_research_data.ipynb and the data is exported as S&P_constituents_statements_ratio.csv. Finally, the labels for the training of the machine learning model are also calculated by computing the difference between stock return and S&P500 return, and it is exported as out_performance.csv.

After that, Macro data such as the SOFR and the COT reports are downloaded from the corresponding websites. The ratios are calculated from the positions disclosed in the COT reports. Then, the micro and macro data are concatenated into one Pandas dataframe, sorted, scaled, and sliced to be ready for machine learning model. The dataset is exported as dataset.csv and being done final_strategies.ipynb.

2.2) Machine learning model & Sub Strategies

The Sklearn package is used for machine learning where the base estimator is RandomForestClassifier() and the adaptive boosting model is AdaBoostClassifier(). GridSearchCV() is used for grid search optimizing the following parameters:

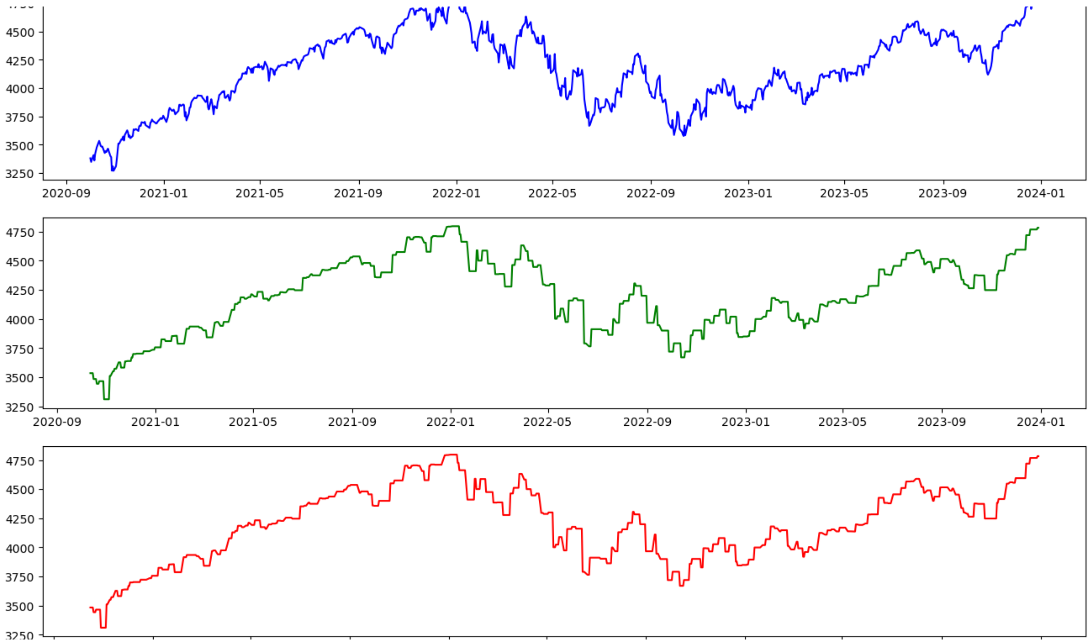
- n_estimators: [10, 20, 30, 40, 50]. The number of estimators (trees).
- 'estimator__max_depth': [None, 1, 2, 3, 4, 5]: The max depth of each tree.
- 'estimator__min_samples_split': [2, 3, 4, 5]: The Min samples required to split a node.

The training set contains 80% of the total dataset and k-fold is set to have 5 splits for the grid search optimization. Time decay weighting mentioned in the literature review will be calculated and input to the model training. After the training, the best estimator is validating again with the test set with the k-fold to get a mean F1 score. Finally, the model is exported as stock_filter_ckp.p.

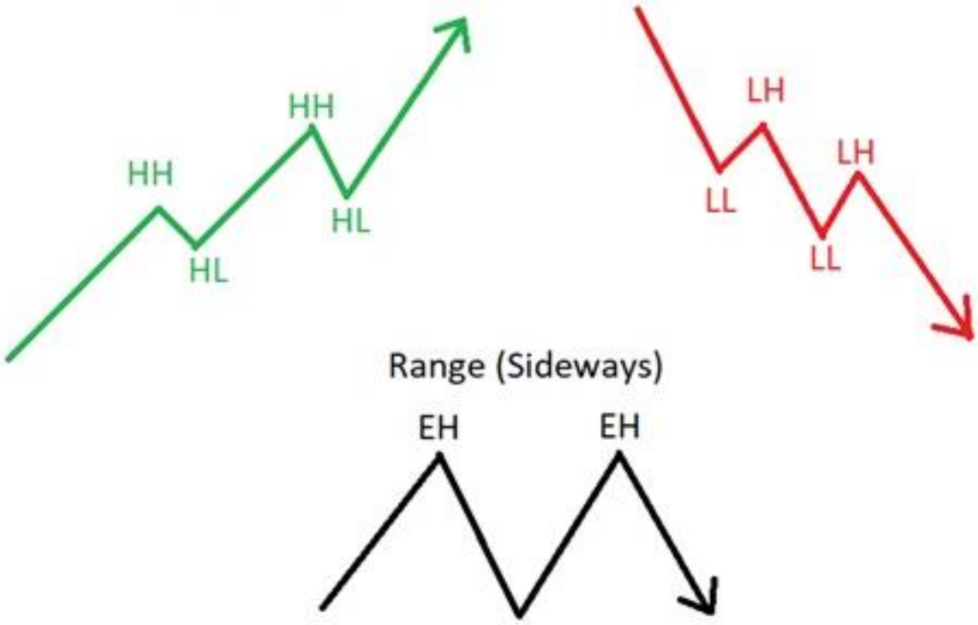
Then, predictions on the out performance of the stocks are inferred by the best model with exported of -1 or 1, and all the stocks with the predictions that are 1 in the test set are shortlisted. Backtesting is done for two sub strategies respectively and the charts of the performance and metrics calculated are plotted.

2.2.1) Smart beta is implemented simply with a buy order at the start and a sell order at the end.

2.2.2) Breakout is implemented with many functions. The market structure is determined by identifying four significant peak and trough points using the "find_pivots()" and "find_local_peak_trough()" functions.



These functions filter out noise, selecting only the most relevant peak and trough points. The four most recent points are then used to determine trends for each data point.



These identified peak and trough points are fed into the "find_trend()" function to analyze both market structure and trend. This function helps determine whether the trend is consolidating or not by identifying three swings. By combining these functions, it becomes possible to determine whether the market is experiencing an uptrend, downtrend, or trading sideways, as shown in the provided figure. The output of the trend analysis is a list of three numbers, either 1 or -1, representing the direction of the swing.

For the logic of the strategy, the strategy first identify a consolidation trend with the find_trend() function, then it record the second and third pivot points from the find_pivots() function since they are going to be the key level for a breakout confirmation. Finally, the strategy generates long or short signals once price breaks above or below key level.

The backtesting results are exported as Backtest_bnh.csv and Backtest_breakout.csv for smart beta and breakout respectively. And were all done in the final_strategies.ipynb.

2.3) Streamlines for real-time data

Real-time data is essential for implementing the strategies in real-time, whether it is pre-live paper trading or live trading. An ideal scenario is when the real-time data pipelines are provided by the APIs, however, the FMP API does not have that for macro data required for the strategies. Therefore, streaming_macro.ipynb and streaming_micro.ipynb are the code done for retrieving real-time macro and micro data, and generate signals. For the SOFR, the CME datamine API is registered to retrieve latest SOFR. But the real-time data from the COT reports must be retrieve through web scraping since the COT reports are not available in the FMP API nor it provides their API. In fact, the request.get() method in the python is blocked by the CFTC website and the data can only be retrieved using Selenium package. After data cleaning and preprocessing, the structured data is exported as macro.csv.

The real-time financial statements for micro data, on the other hand, are provided with the rss feed provided by the FMP API. However, the data is not exactly what the strategies needed since rss feed can only stream the latest financial statements without filtering and avoiding data duplication. Therefore, the program must filter the latest arriving statements first to avoid unwanted data. The program must validate that the new statements are from a stock that is one of the S&P500 constituents, and then retrieve the financial statements of that stock. After that, structure and transform the data with the macro data in macro.csv. Finally, the features are input to the machine learning model and the model will infer whether this stock is going to outperform the S&P500 for the next three months. And then the functions of the sub strategies can be run on the stock predicted by the model and generate the trading signals which will be stored in the signals.csv.

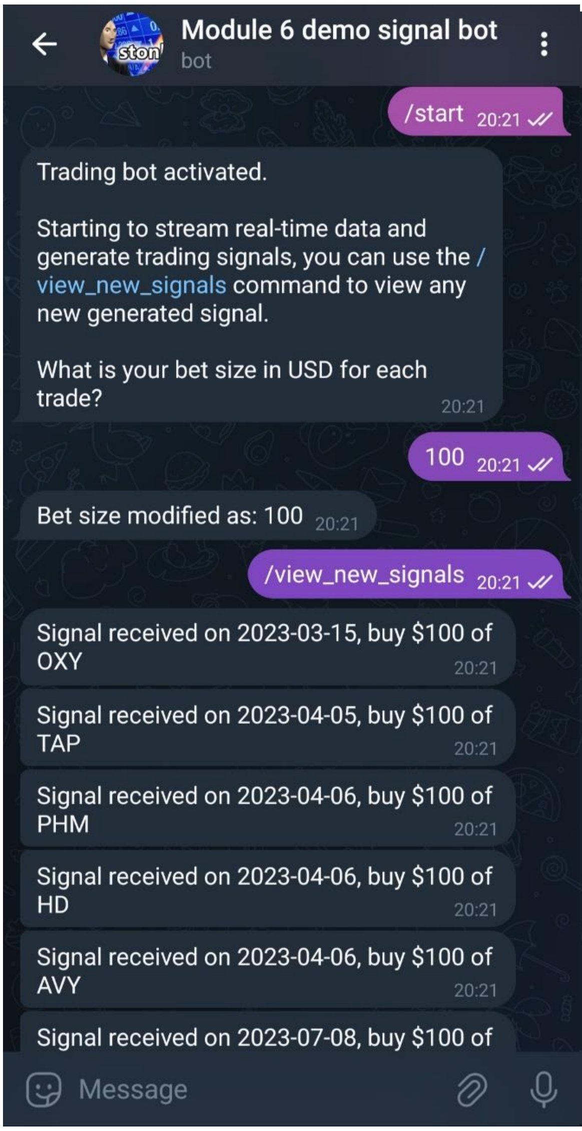
2.4) System for trading and compliance

A system is needed to manage the signals, execute orders, and monitor portfolios for paper trading or live trading. For compliance, the system must log all the placed orders and do trades for regulatory. A position that is too much for a specific stock must be prevented for both risk management and compliance. Moreover, additional functions should be included in the system to allow the firm to conduct KYC on the client efficiently.

Currently, since the strategies is far from being able to profit in live trading, the system is made for stimulating trading. A telegram bot is built to interact with the users and inform them with the new generated signals. The demonstration of the commands and interfere for the bots are shown as follows:

Postgraduate Diploma in Financial Analytics and Algo Trading (Module 6: Algo Trading and Quantitative Investment Strategies)

Student name: Chan Ching Kit
Student number: 20242725



/start 20:21

Trading bot activated.

Starting to stream real-time data and generate trading signals, you can use the /view_new_signals command to view any new generated signal.

What is your bet size in USD for each trade? 20:21

100 20:21

Bet size modified as: 100 20:21

/view_new_signals 20:21

Signal received on 2023-03-15, buy $100 of OXY 20:21

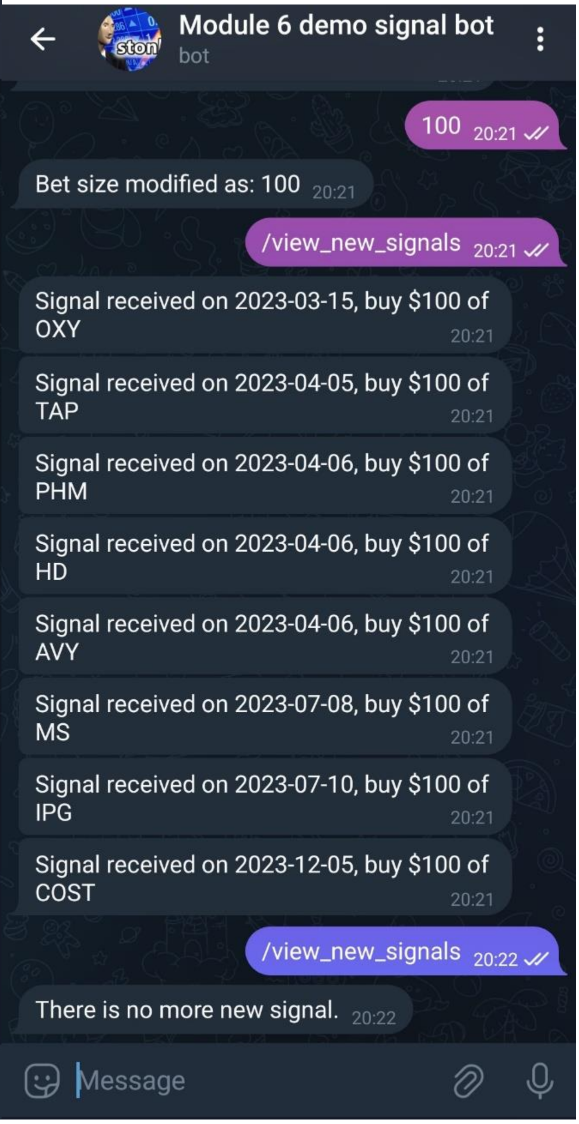
Signal received on 2023-04-05, buy $100 of TAP 20:21

Signal received on 2023-04-06, buy $100 of PHM 20:21

Signal received on 2023-04-06, buy $100 of HD 20:21

Signal received on 2023-04-06, buy $100 of AVY 20:21

Signal received on 2023-07-08, buy $100 of



100 20:21

Bet size modified as: 100 20:21

/view_new_signals 20:21

Signal received on 2023-03-15, buy $100 of OXY 20:21

Signal received on 2023-04-05, buy $100 of TAP 20:21

Signal received on 2023-04-06, buy $100 of PHM 20:21

Signal received on 2023-04-06, buy $100 of HD 20:21

Signal received on 2023-04-06, buy $100 of AVY 20:21

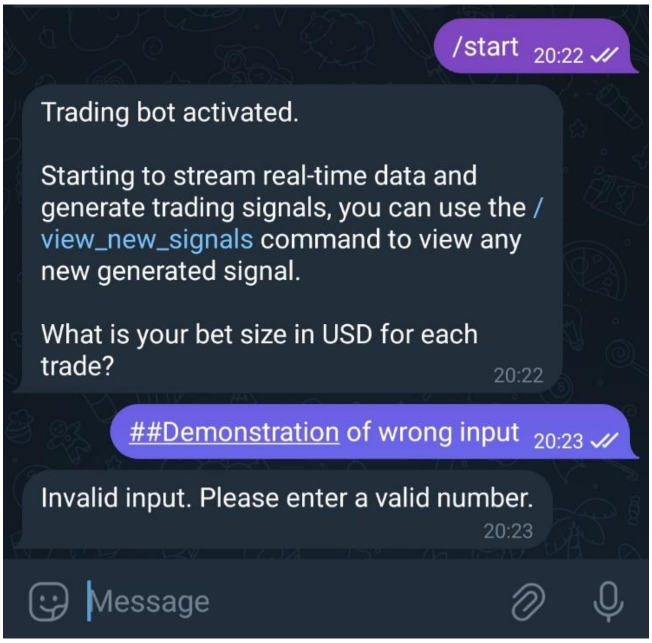
Signal received on 2023-07-08, buy $100 of MS 20:21

Signal received on 2023-07-10, buy $100 of IPG 20:21

Signal received on 2023-12-05, buy $100 of COST 20:21

/view_new_signals 20:22

There is no more new signal. 20:22



/start 20:22

Trading bot activated.

Starting to stream real-time data and generate trading signals, you can use the /view_new_signals command to view any new generated signal.

What is your bet size in USD for each trade? 20:22

##Demonstration of wrong input 20:23

Invalid input. Please enter a valid number. 20:23

The telegram bot can record the user defined bet size and retrieve the new orders updated in the signals.csv and inform the users without repeating old signals that was informed to the users. While at current stage the bot can only inform new signals and all orders are executed manually by the users, the bot is capable for future upgrade including executing trade automatically. However, the automatic executing approach also have its downside of uncertainty as there might be bugs occurring in the program. Nevertheless, the automatic execution can be built as a mode which the users have the option to use. Furthermore, order and position tracking, portfolio monitoring can also be built with the infrastructure of telegram bot. For the KYC of the clients of the funds, the conversation handler functions in the telegram bot can help for easier and faster KYC process where the users can simply input a form in the telegram following the instructions set up in the bot. All the telegram bot features were done in the tg_bot.py.

3) Risk Management and Compliance

Based on the strategies developed, there are risks of over exposure to specific stocks, short squeeze, unlimited drawdown, and inflating gross exposure. Also, there are risks during the operation of the firm that must be considered.

3.1) Procedures and tools

There is a risk for the researchers working in the firm to exploit the secret of the trading strategies, allowing other entities to benefit from the logic of strategies. This can be prevented by signing a non-disclosure agreement when hiring new workers.

However, there is still a chance for the researchers to leak the strategies unintentionally or even intentionally neglecting the non-disclosure agreement. The strategy development structure ensures that during the development of the strategies, there is no researcher that can know all the logic and parameters for the strategies that the firm has developed. So, the intellectual properties of the firm can be protected beside non-disclosure agreements.

For the exposure risks occurring in the strategies, while the smart beta strategy might be fine, a position limit can be established for the breakout strategy. The gross exposure can be capped at a certain level where the algorithm will rebalance the portfolio, taking profit or stoploss to generate funds for new orders instead of adding new funds for new orders. Also, the betting size can be optimized by referencing the data from the backtest and the Kelly criterion. The breakout strategy can be modified as taking only long position to eliminate the risk of short squeeze. Finally, although the downside is limited to the breakout strategy as it is protected with the stoploss order, the smart beta strategy is not protected. Even though the smart beta is only taking long trade and there is impossible for it to lose more than the fund invested as long as no leverage is being used, it is better to set a stoploss order to avoid black swan event. Still, the stoploss can be set with a higher percentage, can be referencing from the VaR value from the backtest, to avoid the objective of buy and hold.

3.2) Regulatory and Compensation Discussion

When operating a trading firm, it is essential to consider regulatory frameworks to minimize potential violation of the regulation. While the regulatory measures for the stock market have a longstanding history and exhibit consistency, the regulation concerning AI is currently under discussion and may be established in the future due to the rapid advancements in AI technology.

3.2.1) Due to the algorithm's tendency to close positions within a day, if it consistently adheres to take profit and stop loss levels, it could potentially violate the Pattern Day Trading (PDT) rule[2]. To engage in day trading, a margin account must maintain a balance of at least $25,000.

3.2.2) Trading firms are required to adhere to the Anti-Money Laundering (AML) rule[3], which aims to prevent and identify activities associated with money laundering and terrorist financing due to the substantial volume of trades conducted by these firms. In their operations, firms should perform Know Your Customer (KYC) and Customer Due Diligence (CDD) processes to verify customers' identities and the sources of their funds. Continuous monitoring of transactions is necessary, and any suspicious activities must be reported.

3.2.3) The Securities and Exchange Commission (SEC) enforces the Anti-Fraud Provisions of the Investment Advisers Act, which prohibits investment advisers from participating in fraudulent activities, including market manipulation, or providing false or misleading information to clients. To ensure accuracy and facilitate future investigations, the firm should implement double checks on all documents or statements provided to customers.

3.2.4) Although there are no explicit regulations specifically addressing the use of AI in trading, the SEC has issued guidance regarding the application of AI and machine learning in the financial industry. This guidance emphasizes the

[2] https://www.tradingsim.com/blog/pattern-day-trading-rule

[3] https://www.sec.gov/about/laws/secrulesregs

importance of transparency, accountability, and robust risk management when employing AI in trading activities. To better align with potential future regulations, the firm should consider utilizing machine learning models that are more interpretable, such as decision trees and random forests.

Data Compilation and Findings

All the important data in the project are exported to csv files, including S&P_historical_constituents.csv, signals.csv, macro.csv, etc. And the performance and the metrics from the strategies backtesting are plotted into charts, shown as follows:

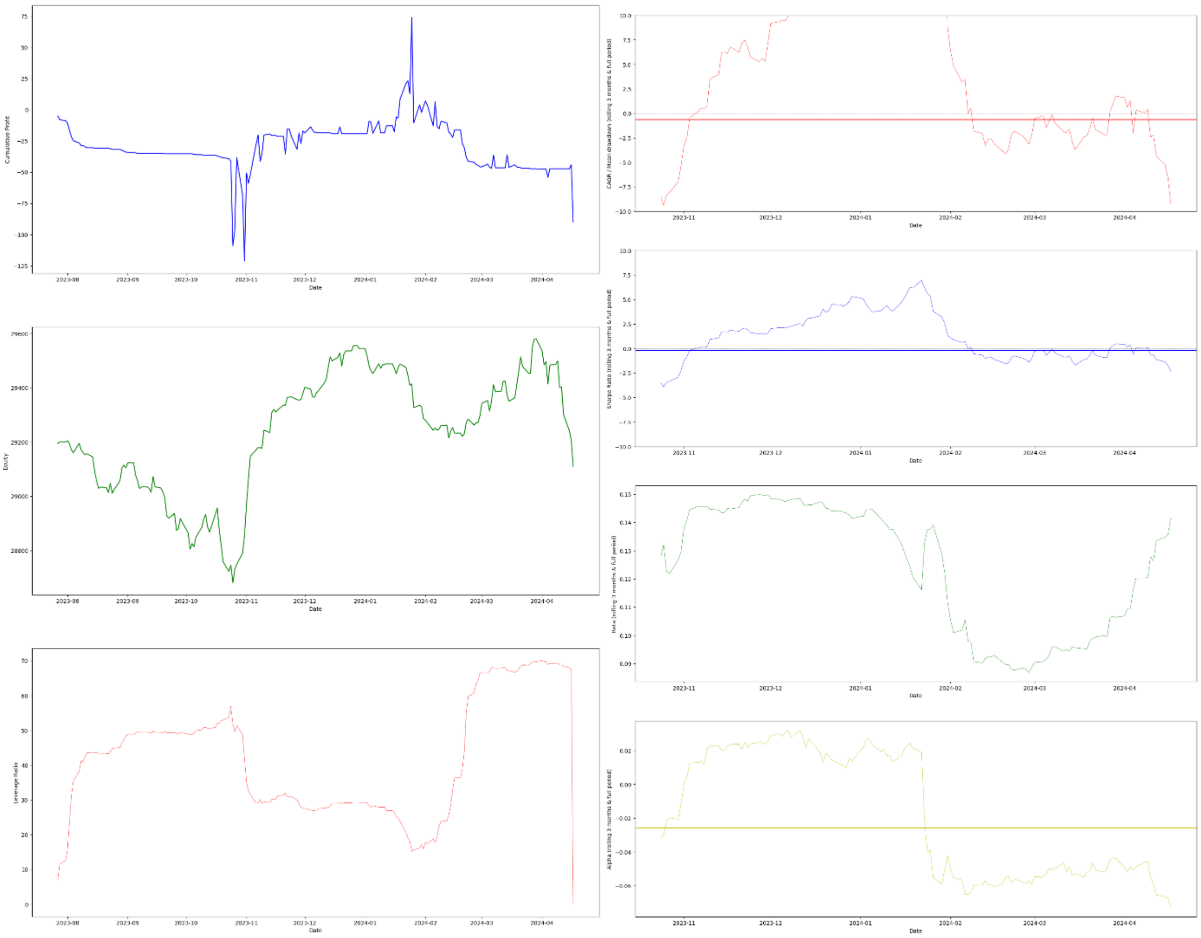


Figure 1 Smart Beta strategy.

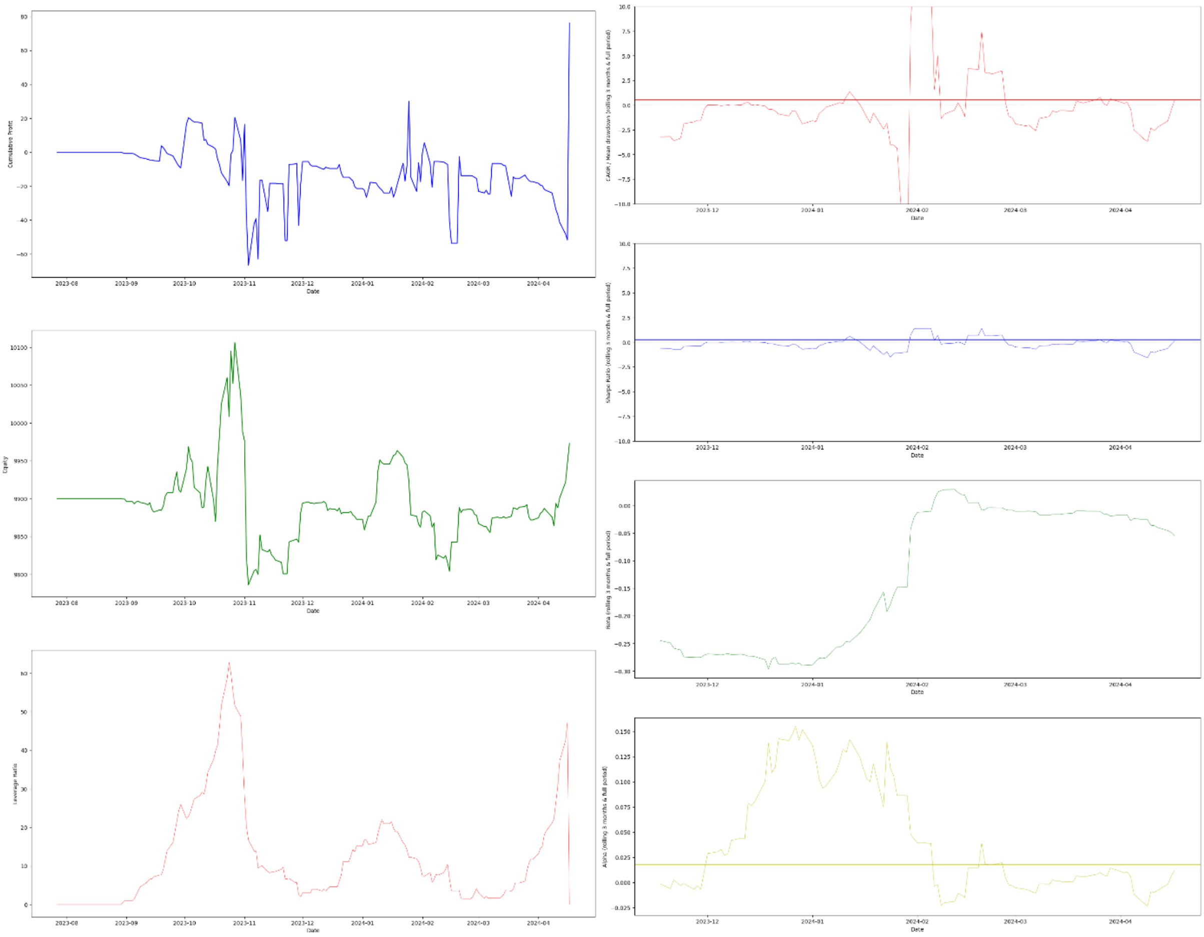
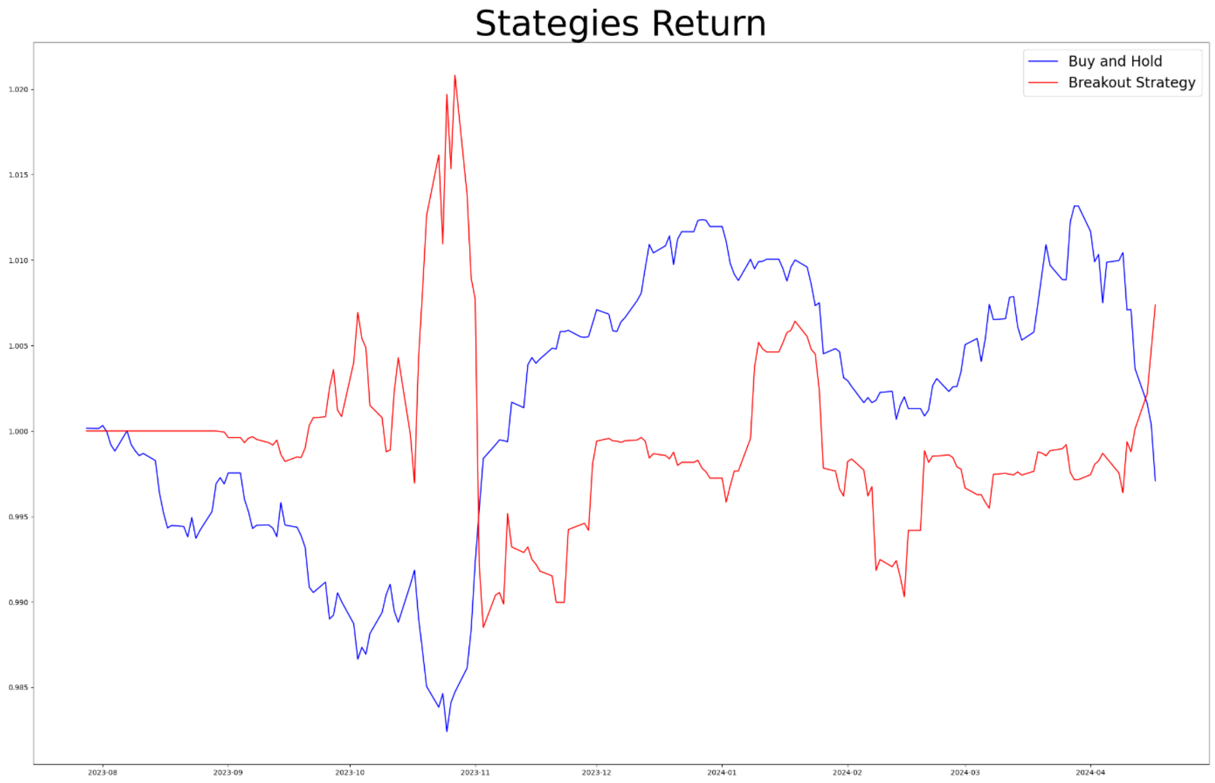


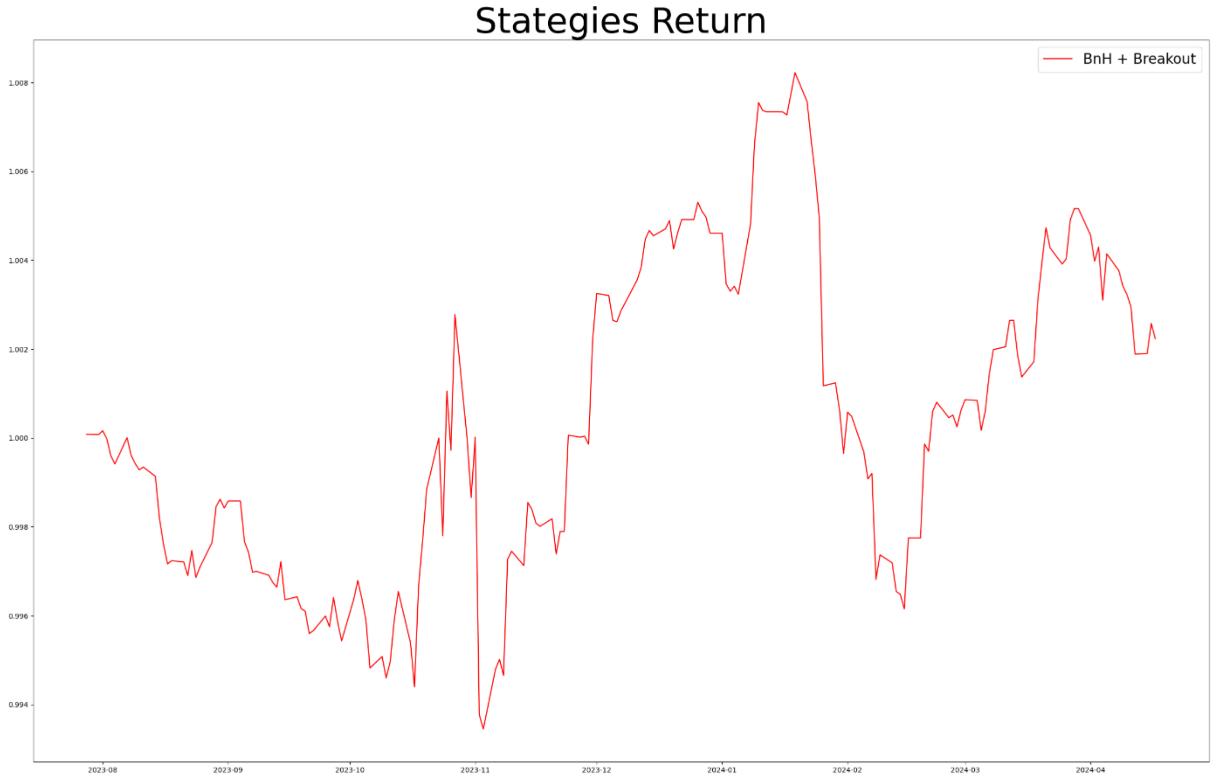
Figure 2 Breakout strategy.

The charts on the left plot the details of the strategies including cumulative profits, equity curve, leverage ratio. Both strategies appear to be inconsistent in their cumulative profits, with the breakout strategy having more fluctuations in it due to more trade being done. The equity curve shows that both strategies are not consistently profitable. In fact, the smart beta strategy results in a loss at the end and the breakout strategy results in a slight profit, and it is likely due to randomness. Also, the range of the equity is not huge, maybe due to the fund not being utilized fully most of the time. It makes sense because the signals are generated only when companies release their financial statements, and the companies release it quarterly which created gap periods of little to no signal generated. The leverage ratio tends to spike up and plunge down around the period of releasing of quarterly statements.

The charts on the right plotted the performance metric (rolling 3 months and full period) based on the return of the strategies, including CAGR/ Mean drawdown, Sharpe ratio, Beta, and Alpha. Both strategies have their CAGR/ Mean drawdown around 0, indicating the strategies are not profitable but also not having a lot of loss. There are periods where their profitability spikes up, but they were not sustainable. The Sharpe ratio is also around 0 for both strategies, indicating holding risk-free asset might be better than the holding the portfolio created by the signals. Both strategies have low beta whereas breakout strategy even have negative beta for nearly half of the period, indicating they are not sensitive to S&P500 return. The alpha for breakout strategy is close to 0 and for smart beta it is slightly negative, indicating they are performing worse than their expected return based on the risk they carried and the market environment.



However, when plotting together the two strategies. It is easy to notice that one of the strategies is having huge profits when the other strategy is experiencing huge losses. This result has created the wonder to combine both strategies to hedge against each other.



The chart above shows the equity curve of the combination of the strategies with the equal allocation. Although it seems better than both strategies where the combination shows a slight uptrend, the performance is still very inconsistent and not very profitable.

Discussion, Analysis and Implications

Overall, the analysis suggests that the evaluated strategies, despite their utilization of a machine learning model and sub-strategies, do not yield consistent profitability. The equity curves display fluctuations, and the performance metrics indicate lackluster results.
It appears that the strategies developed in the project do not demonstrate consistent profitability or outperformance compared to the market. The inconsistent cumulative profits, equity curves, and performance metrics indicate that the strategies may not be effective in generating sustained returns. Factors such as infrequent signal generation and potential randomness in profitability contribute to the lackluster results. Further refinement or exploration of alternative strategies may be necessary to improve the performance and achieve more desirable outcomes. Therefore, the strategies should not be put into live trading at current stages, and it would be a better choice to hold risk-free assets instead of trading with the strategies researched in the projects.

To deal with the problem of inconsistent profitability and high volatility, several changes can be made to improve the strategies. For the machine learning model:

- Consider incorporating additional features or risk factors into the machine learning model to improve its performance.
- Try out other statistical and mathematical preprocessing on the dataset.
- Explore other ways to adjust the model's parameters or weighting scheme to optimize its performance.
- Consider combining the machine learning model with other strategies or risk factors to improve overall performance.

For the sub strategies:

- Consider incorporating additional risk management techniques to reduce the fluctuations and improve overall performance.
- Explore ways to reduce the leverage during periods of high market volatility to minimize losses.
- Consider incorporating other market indicators to improve the strategy's performance.
- Consider re-evaluating the underlying assumptions and parameters of the strategies to identify potential issues.
- Consider combining the Smart Beta strategy with other strategies, and experiment with different allocation to improve overall performance.

Conclusion

The algo is mimicking a top down and bottom-up investment approach. With the machine learning model acting as a fundamental analyst and the rule-based strategy generating entry and exit signals. The project aims to automate an all-rounded analysis, which is what most stock analysts have been doing. Differs with embedding a machine learning model in the strategy, this approach resulting in a more dynamic algo while preserving scalability.

The machine learning-based trading strategy presented in this paper demonstrates the potential for using artificial intelligence to inform investment decisions, even build a trading firm around it. However, the results of our backtesting also highlight several limitations and potential areas for improvement.

Firstly, while the machine learning model was able to generate signals with potential for profit, the overall performance of the strategy was not consistently profitable, and the volatility of the portfolio was a major concern. This suggests that further work is needed to develop more robust and sustainable trading strategies.

Secondly, the use of backward-looking financial statements as inputs to the model may limit its ability to capture forward-looking information and adapt to changing market conditions. Future work could explore the use of alternative data sources, such as sentiment analysis or real-time news feeds, to improve the model's predictive power.

Thirdly, the increasing exposure of the portfolio over time may also necessitate the development of more sophisticated risk management techniques in position sizing or stop losses, to mitigate potential losses.

Finally, the relatively short backtesting period may not be representative of the strategy's performance over a longer period. Future work could explore the use of longer backtesting periods or walk-forward optimization to evaluate the strategy's performance over a more comprehensive period.

Despite these limitations, the results presented in this paper demonstrate the potential for machine learning-based trading strategies to provide valuable insights and predictions for investment decisions. Further research and development are needed to improve the robustness and sustainability of these strategies, and to overcome the challenges posed by limited data and volatility. Nevertheless, the architecture established is easy for scaling and further development. And the mandate and objectives of the hedge fund was achieved in the project.

Limitations of Research

The first limitation is the low accuracy of the model, which affects all the sub strategies as they are bind together. It is likely due to not having enough data for training. There isn't enough data available for the machine learning model training and this can be a significant obstacle, since machine learning models typically require large amounts of data to accurately learn and make predictions.

The second limitation is that the financial statements are backward-looking in nature. This can make it difficult to use them for predictive purposes, as they are focused on past performance rather than future outcomes.

Moreover, the backtesting period is too short, it may not be representative of the model's true performance over time. It is due to 80% of the data is taken for the machine learning training which means only 20% of the data is available for testing. 80% cannot be used for testing since the model will likely to generate unrealistically good result as it is optimized with the training data. One workaround is to split the data with more percentage in testing set, but then it will affect the training of the model.

Lastly, the work from the operation segment in this project is at most a stimulation of real trading and it serves for the discovery of the issues that will be encountered. Nevertheless, with the challenges encountered, a well-rounded procedure is established as follow and will be done if resources and time allow:

- Communicate with other segments and retrieve all the data needed.
- Clean and preprocess all the data, export to database in a structured manner.
- Build the program for machine learning and sub strategies backtesting.
- Tailor-made the system to stream real-time data and stimulate trading, and the extra features for risk management and compliance related work.
- Paper trade with strategies with promising backtesting results to test on things like slippage and connectivity.
- Monitor the paper trade result for a predefined period that is long enough for proper risk management, build the extra programs needed for the live trading.
- Start live trading with small amounts of capital and keep monitoring, add funds slowly to the algorithm if the result is well and under fully control.

References

CME datamine.
Financial Modeling Prep API.
Prado, M. L. (2018). In *Advances in Financial Machine Learning.*