International Journal of xxxxxx
Vol. x, No. x, (2019), pp. xx-xx

Application of algorithmic trading strategies for retail investors.

Mikhail Shishlenin, Ganesh Harke, Suresh K Koppisetti
WorldQuant University MScFE
shishlenin@gmail.com, ganeshah1711@gmail.com, suresh.koppisetti@gmail.com

Abstract

Retail investors are an essential part of the financial markets and they provide stability to the markets by providing liquidity and helping in price discovery. For markets to function efficiently, a healthy retail investor landscape is essential. However, with the advent of HFT and institutional algo traders, the retail investors are left on the sidelines and participating in the financial markets through some ETF. This paper is an exploration of the application of algorithmic techniques at a retail investor's scale. Using simple quality and value factor models to moderately sophisticated machine learning techniques like Random Forest and AdaBoost, we have explored if a retail investor can participate in the capital markets and earn better returns than by investing in ETFs. Given that factor investing and various rules-based strategies have previously been studied in academia, we fill the gap in the literature by providing our own variables to each factor as well as testing performance across two geographical regions. The empirical analysis performed in this thesis suggests smart beta algorithmic trading or AI based strategies outperform on risk adjusted basis. We therefore conclude that retail investor can leverage on algorithmic techniques and AI based strategies such as smart beta or factor investing to create medium to long term trading models.

Keywords: *algorithmic trading, retail investors, medium-term trading models, long-term trading models, machine learning, smart beta, factor investing, random forest, ada boost, dollar bars, triple barrier method, meta-labelling, S&P 500, MOEX, QuantConnect*

1. Introduction

In 2019, most of the trading and investment decisions were performed by automated algorithms. With institutional players dominating the field. However, during the same period, there is an influx of retail investors in the market partly due to the rise of discount brokerages [1]. Moreover, "... 67% of millennials saw recommendations made by artificial intelligence as being a basic part of any investment platform…", and they are more willing to actively react on market volatility in comparison to older generations [2].

The goal of this work is to explore active investment opportunities of the retail investor and build Python-based application to execute the medium- and long-term trading strategies. For the purpose of this work, several approaches should be researched for an inference regarding their profitability and practicability for the retail investments (say, 0.2-5M USD AUM). Among strategies under consideration are: mean-reversion, swing trading, tactical asset allocation, dynamic position sizing, etc. Another important part of the successful implementation of the above mentioned strategies is the proper preparation of incoming data (feature), which should include application of financial time-series

ISSN: XXXX-XXXX IJXX

specific tools for model validation and backtesting [3]. In order to have low trading fees and high liquidity the scope of current research is mostly focused on US exchange traded stocks.

2. Literature Review

There is a lot of research conducted on algorithmic trading in many areas but most of this research was done for institutional investors and portfolio managers such as Hedge Funds, Mutual funds and Investment banks. Also, there is a lot of research that was done on retail investors but most of it on behavioral analysis and individual investor as market participant with limited resources. There is very little research on Artificial Intelligence (AI) as it pertains to a retail investor.

Institutional investors with their access to cutting edge innovations in Machine Learning and Deep Learning and their application in algorithmic trading, react to market information in a fraction of a second. On the other hand, a regular retail traders are always front run by these sophisticated algorithms. This loss can be seen in terms of ex-post trade cost [4].

An individual investor is limited to doing manual analysis and research by searching for information on the web, reading companies 10K filing which is very time consuming and more often than not, by the time the individual investor reacts to the information he/she had been researching, the effects of this information, should it be material, is already reflected in the market price of stock. On the other hand, the institutional investors are employing techniques such as Natural language processing (NLP), Sentiment Analysis etc., to react to new information at lightning speed without too many manual interventions.

Retail Investors are prone to biases [5], [6], [7], [8] to name a few: Mental Accounting, Anchoring, Gambler's Fallacy, Availability, Loss Aversion, Regret Aversion, Representativeness, Overconfidence and Optimism, Herding and Disposition effect. Thus [9] found that retail investors mostly perform technical analysis, which is associated with greater portfolio concentration, more turnover, less betting on trends, more options trading, a higher ratio of non-systematic risk to total risk, lower gross and net returns, and lower risk-adjusted returns.

Individual investors follow momentum and contrarian strategies, thus [10] claim individual investors decrease volatility and provides liquidity. Individual investors show trading skills improvement after period of overtrading, whereas according to [11] disposition effect still remains. Moreover, they could time the market [12], and are familiar with and exploit market movements attributed to the political environment in Washington; some wealthy groups of individual investors may consistently outperform the market [13]. Although, cognitive limitations were found with tasks demanding greater mathematical reasoning ability which is more relevant to performance [14].

One more consideration to mention is the predatory algorithms of high frequency technology (HFT) traders [15], who exploit traders' behavior and infrastructure features to gain on others. In order to decrease such patterns investors could benefit from processing raw tick data and receive similar informational advantages.

The democratization of backtesting frameworks and drop in the transaction costs at all major brokerage houses are a blessing to the retail investor in leveling the field to compete with institutional players.

3. Model Design

Investment return is driven by multiple factors and strategies. Below is a high level view of drivers of investment returns [16].

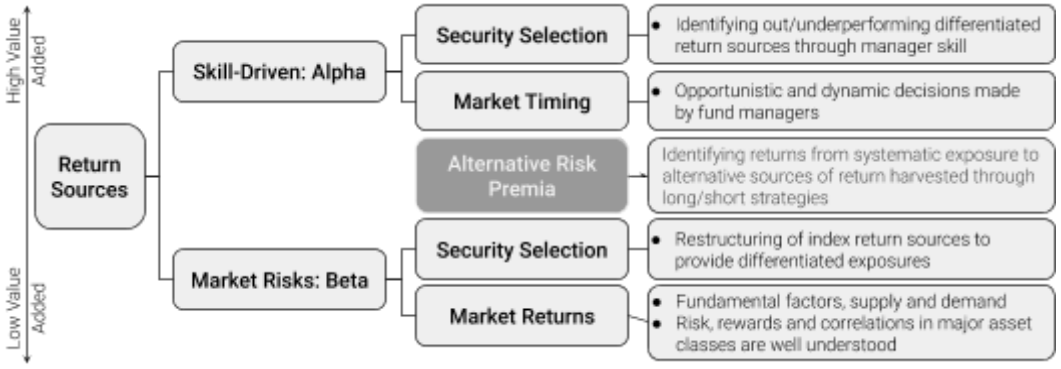

Figure 1. Drivers of Investment Returns

In this research, we are picking smart beta strategies to assess whether individual investors can generate alpha using different factors.

3.1. Smart Beta

Smart beta refers to an enhanced indexing strategy that seeks to exploit certain performance factors in an attempt to outperform a benchmark index. Smart betas are non-cap-weighted index strategies based on transparent quantitative methodologies. Deviating from cap weighting in a systematic way helps address the flaws of cap-weighted indexing. Investors choosing among smart betas should consider if they are more sensitive to volatility or underperforming the S&P 500 [17].

Alternative equity indexes are likely to outperform traditional cap-weighted indexes over the long term, research results show that such smart beta strategies are exposed to several types of risk, including systematic risk (e.g., factor tilts), specific risk (related to the assumptions and inputs of a strategy), and relative risk (i.e., the risk of potentially severe underperformance) compared to cap-weighted indexes that can last for extended periods of time [18]. Rather than relying solely on market exposure to determine a stock's performance relative to its index, smart beta strategies allocate and rebalance portfolio holdings by relying on one or more factors. A factor is simply an attribute that might help to drive risk or returns, such as quality or size [19].

3.1.1 Trading Signal and Trade Timing

Trade timing means generating trade signals based on price trends. The two main trading strategies are momentum and mean reversion. Mean Reversion in finance is an assumption that the price of a stock tends to move towards the long term mean or average price level. Most of the people in economics believe that what goes up will eventually come down [20].

A simple approach to trade “mean reversion” to take a long position when there is a huge drop in stock price and short the stock when there is a sharp rise in stock price. It examined the possibility of a profitable mean reversion trading strategy by creating and backtesting a mean-reverting trading algorithm [18].

There are different ways to trade mean reversion based on technical indicators like Bollinger Bands, Relative Strength Index (RSI), Standard Deviation and so on;

fundamental indicators of stock; economic indicators such as GDP, inflation, interest rates, etc.; sentiment indicators. As an initial choice for our strategy, we fill rank the stocks and take long positions when stock is in the top 20 and go short as it loses position in top 20 stocks. Further, we will adjust our approach based on the performance indicators.

3.1.2 Research Design & Methodology:

We are going to consider multi-factor smart beta strategy since it will give us methodological consistency, implementation efficiency and diversification. We are going to test below strategies:

1. Portfolio management based on "*Value Smart Beta*" strategy using fundamental factors such as dividend yield, PE, PB and PCF.
2. Portfolio management based on "*Quality Smart Beta*" strategy using fundamentals such as ROA, ROE and Debt to equity ratio.
3. Active portfolio strategies based on machine learning which comprise different fundamental factors.
4. Active portfolio strategies based on high-frequency tick data processed into custom bars which will be fed as input to our machine learning models.

Our tradable universe comprises the most prominent class of common stock of all listed US equities. For a holding period, at the beginning of each trading month, our algorithm runs the strategy on the entire tradable universe and selects configured numbers of stocks. Next, it would take positions in these stocks based on the top *N* number OF stocks for which long signals generated by our algorithm. Since we are using a long-only strategy, it is sensitive to market risk, and will make it fall short of pure alpha. Although the term of rebalancing (daily, weekly, monthly, quarterly, or other) will depend on profitability for the retail investor.

Once technical and fundamental data is available to strategy, we will pass data to different strategies to generate the signals and take the position at the start of every month. For machine learning algorithms, we will train our model on historical data and update the model every month before updating the portfolio. We choose S&P 500 as our benchmark[1] . However, due to complicated data access and computational limitations for the processing of tick data we select 16 blue chips constituents of MOEX index for the algorithm building and analysis. We have followed equal weighting schema for portfolio creation and taking position in individual stock and leveraged to Quantconnect model.

3.2. Factors

Factors are the set of characteristics or fundamentals of securities that explains performance and risk of the security. As per Capital Asset Pricing Model (CAPM) developed in the 1960's, one single factor explains the performance, which is its exposure to the market portfolio, otherwise known as Beta. In this thesis, we are exploring value and quality factors against benchmark S&P 500.

3.2.1 Value:

Value strategy has a long history from the Graham and Dodd era. The strategy aims at identifying undervalued securities based on their price-to-fundamental ratios. Different types of ratios are examined extensively by academia such as price-to-cash flows, price-to-earnings, price-to-book, as formalized by Fama and French [21].

[1] S&P 500 is a stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the United States.

The economic intuition of the Value factor requires the investor to assume that a relatively low (high) fundamental value indicates that the asset is undervalued (overvalued), hence different measures of price-to-fundamentals are commonly used to generate the signal. The investor will typically invest in a portfolio of undervalued assets in relation to the market, expecting the portfolio to outperform the corresponding index until levels are in line with the rest of the market. The expected effect is therefore exponentially decreasing as it returns to market standards. This suggests that most excess returns occur early in the business cycle [26].

In this thesis, we are considering four different value smart beta factors to generate the signals. We are considering price-to-earning (PE), price-to-book (PB), price-to-cashflow (PCF) and dividend yield (DY) ratios (Details for all these variables are in Table 1). We are calculating Z-score (standardized score) for each variable and then calculating weighted average of standardized score for each security. These calculations will happen at the start of every month hence score and statistics are running statistics (Appendix A explains details of standardized scores). The value score for security *i* is calculated as follows:

$$VS_i = (-1) * [0.25 * Z_{PB_i} + 0.25 * Z_{PE_i} + 0.25 * Z_{PCF_i}] + 0.25 * Z_{DY_i} \quad (1)$$

Where:

VS_i : Value Score of security *i.*

Z_{PB_i} : standardized score of price-to-book ratio of security *i.*

Z_{PE_i} : standardized score of price-to-earning ratio of security *i.*

Z_{PCF_i} : standardized score of price-to-cash flow ratio of security *i.*

Z_{DY_i} : standardized score of dividend yield ratio of security *i.*

First three standardized variables by -1 in order to obtain a higher (lower) score for a lower (higher) fundamental to price ratio. The top 50 value scores (rank based) are selected and weighted according to our weighting scheme.

3.2.2 Quality:

Different academic studies and research has proved that quality companies have high excess returns.Also studies show that cash flow fundamentals steer stock prices more than macroeconomic variables, meaning a well-run firm can gain a competitive advantage through careful capital management. Quality stocks tend to perform better during bad times because if macroeconomic conditions start to deteriorate, investors will become risk-averse and start investing in stocks with sound capital management. This will result in a "*flight-to-quality*" effect which means push up the value of quality stocks [26].

There are different theories to identify the quality stocks, one of them is well known from Piotroski. It generates an F-Score, which determined the financial strength of a company by the sum of nine binary variables. There are more simple approaches to define quality as well such as firms with high gross profitability earned returns in excess to the market benchmark over longer periods.

In this thesis, we are going to utilize a simpler one dimensional approach to identify quality of stocks. We have constructed quality smart beta by combining various fundamentals of stocks. Instead of assigning binary value to each quality, we will use standardized Z-Score of each variable to identify the quality of stocks. Historical research has proved that low variability in earning-per-share results into high quality stock. We have incorporated such variables which gives indicators of quality.

We define the Quality Smart beta as weighted average sum of standardized scores of five quality factors. Five quality factors that we have selected includes debt-to-equity ratio to capture the leverage factor; return-on-equity (ROE), return-on-assets (ROA), and operating cash flow (CFO) capture the profitability factor complemented by earnings variability to capture earnings quality (All these factors/variables are explained in Table 1). Quality score for a security, *i*, is calculated as below:

$$QS_i = (-Z_{DE_i} - Z_{EPSVar_i} + Z_{ROA_i} + Z_{ROE_i} + Z_{CFO_i})/5 \quad (2)$$

where:

QS_i : Quality Score of security *i*

Z_{DE_i} : Standardized score of debt-to-equity ratio of security *i*

Z_{EPSVar_i} : Standardized score of earning-per-share variability of security *i*

Z_{ROA_i} : Standardized score of return-on-asset of security *i*

Z_{ROE_i} : Standardized score of return-on-equity of security *i*

Z_{CFO_i} : Standardized score of operating cash flow of security *i*

Top 50 scores are selected and weighted according to our weighting scheme.

3.2.3 Quality and Value Using Machine Learning:

In this smart beta strategy, we have combined all the variables of quality smart beta and value smart beta strategy with some additional variables to generate the signals for each stock. We have combined all these variables and processed with three different machine learning algorithms. To train the model, we added signal long and short using Simple Moving Average (SMA) method. We identified when 20 day SMA crosses 60 day SMA from above go long and when 20 day SMA crosses 60 day SMA from below go short. Variables that are used to train models are explained in Table 1.

As mentioned earlier, three machine learning algorithms used in thesis are as below:

1. Random Forest
2. Ada Boost

For all three models, we have used sklearn python library which is considered as de facto for machine learning in Finance.

Below is a detailed description of fundamental factors used in this thesis. We have leveraged QuantConnect to provide all fundamental data which is provided by Mornigstar.

Table 1: Fundamental Factors or Variables

Factor	Formula	Description
Price-to-Earning (PE)	$PE = \frac{Market\ Price\ of\ Stock}{Earning\ per\ stock\ unit}$	Determine whether shares are correctly valued in relation to one another
Price-to-Book (PB)	$PB = \frac{Market\ Price\ of\ Stock}{Book\ value\ of\ stock}$	Used to compare a company's current market value to its book value
Price-to-CashFlow (PCF)	$PCF = \frac{Market\ Price\ of\ Stock}{Operating\ CashFlow}$	Used to compare the value of stock to operating cash flow.
Dividend Yield (DY)	$DY = \frac{Stock\ Dividend}{Price\ of\ stock}$	Used to determine cash flows generated by stock

Cash Flows from Operation (CFO)	$CFO = EBIT + Depreciation - taxes + \Delta working\ capital$	Used to determine cash flows generated from operations
Return on Asset (ROA)	$ROA = \frac{Net\ Income}{Total\ Asset}$	Reflects by percentage how profitable a company's assets are in generating revenue
Return on Equity (ROE)	$ROE = \frac{Net\ Income}{Shareholders\ Equity}$	Used to measure how well a company uses investments to generate earnings growth
Debt to Equity (DE)	$DE = \frac{Total\ Debt}{Total\ Equity}$	Used to determine total leverage of company
Earning per Share Var (EPS)	$EPSVar = \sigma_{\Delta EPSGrowth;\ T;T-1}$	Used to determine earning stability or growth rate
Sales to Price (SP)	$SP = \frac{Sales}{Stock\ Price}$	Used to determine the value of a stock relative to its past business performance
Market Cap (MCAP)	$MCAP = Market\ price * No.\ of\ shares\ outstanding$	Reflects how much money is raised and the size of listed companies

3.3 Advances in ML based on Dollar Bars and Triple Barrier Method

Building trading strategies by processing raw tick data were performed by hedge funds since late 80s (Medallion, Renaissance, etc.) Following research is based on the DollarBars and Triple Barrier Method, which were beneficial [3] for many years. According to Triple Barrier Method and Dollar Bars [3] as well as the proper implementation [27], the raw tick market data is processed into the 'DollarBars', which are sampled when specified amount of asset is traded, in contrast to ordinary 'TimeBars' - sampled by time (daily, hourly, etc.) In addition, tick i.e. traded equity price should be adjusted for dividends paid. By relying on DollarBars the data is sampled more frequently in the periods of intense trading (high volume). That leads to timing the market changes as well as improving data statistical properties for the following ML.

3.3.1. Features

Based on prepared bars, we will produce a set of features: volatility, serial correlation, moving average, etc. for the further alpha extraction. Apart from that, we will need a labelling features to run ML algorithm. Thus, we will, based on [3] suggestion to use meta-labelling (including triple-barrier labelling) for the trade decision making (see Appendix B).

For this research we applied triple barrier method for making yes-or-no confirmation decision to bet (meta-labelling), when decision on the side was already done by simple moving average method. We run algorithm with set of target return parameters: 0.1-3%, and with set of vertical barrier length: 1-50 days. In addition, SMA was based on 20/50 and 50/200 windows, although didn't show much difference.

To address binary classification problem, for each of the parameter the RF algorithm was trained to know train/test efficiency with precision, recall and f1 indicators. The future development of this work may incorporate the calculation of probabilities to assess the size of the bet based on [23] the recommendation. In addition, portfolio construction/rebalancing for retail investor may be determined by trading costs QC model, so far optimal frequency should be middle or long term. As an approach, waiting period can be introduced for better performance [24], in addition NN algorithms can also provide a solution for the better timing decision [25] with non-linier multi-period trade schedule optimization to determine optimal trading strategies.

We consider the implementation of ML - Random Forest Classifier to learn trend following trading strategy based on features and meta-labelling to decide on the trade. In addition, the model should be re-learned on ongoing recent data to consider latest changes. When applying the features one should take into consideration risks of overfitting. Thus, CV and test measures (precision, recall, f1) should be calculated. One more thing to consider in the future development of current research, information leakage while cross validation. According to [3] that could be prevented by Purged-K-Fold CV.

3.3.2. Implementation Approach

To build this algorithm will require to breakdown the task into several parts.

1. Firstly, we will need to download a 10 years' tick data for 16 blue chips of MOEX companies (as they were in 2014, to reduce survivorship bias) with the help of separate python script (data was provided by one of Russian brokerage companies).
2. Secondly, that data was adjusted for the dividend paid backward.
3. Thirdly, the data was processed into DollarBars (which are actually RubleBars for this research), sampled each time a predetermined amount of asset was traded. In addition, mentioned above technical indicators were built based on this bars.
4. Fourthly, a triple barrier method was applied to generate a set of labels for further training of RF.
5. Lastly, QC is loaded with preprocessed data for the periodic ML RF training on historical values. Then ongoing data is used to create alpha signals and properly invest into assets.

Starting with a sort of equally-weighted portfolio, in the future could be improved this part of the solution implementing more advanced bet sizing, i.e. based on maximum predicted probability [3].

3.4 Backtesting Framework

From the point of view of the retail investor, which could potentially be a person or small group, it is beneficial to rely on framework for reducing mistakes and get time-to-market efficiency. As the students of quant program, for the purpose of this research we are focusing on the QuantConnect (QC) platform which requires a bit advanced skill set, we realize that many individual investors would prefer easier choices like Tradingview or Quantopian.

We choose QuantConnect as a major platform for this research due to several reasons. Firstly, QC provides raw tick data that we plan to process. Secondly, QC fully supports python environment with open-source LEAN engine as a core of the platform. Thirdly, QC considers brokerage fees and etc. payments, moreover it contains fee scheme for Interactive Brokers, which is popular among retail investors. Lastly, as some other platforms, QC provides ready-to-use framework (see. Figure 2), which could be customizable by introducing users' python classes. However, it's worth mentioning that QC will probably requires more skills for retail investors to use it.

As it stated in documentation [22] the figure below depicts how QC framework works.

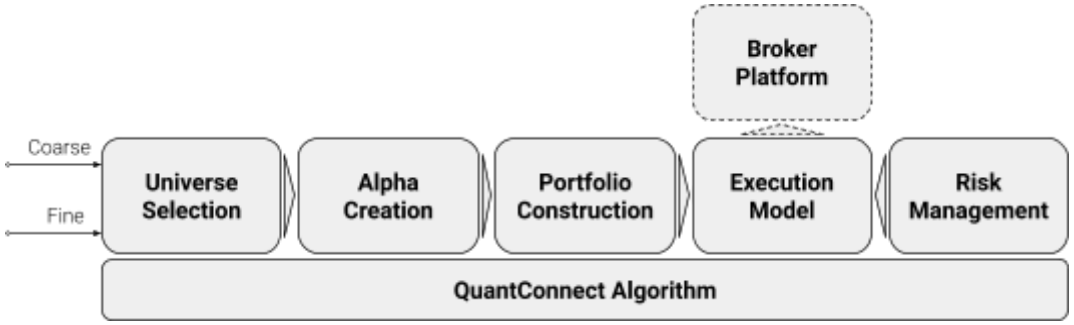

Figure 2. QuantConnect Algorithm Framework Architecture

We have leveraged QC for trade execution and risk management. We have used *"ImmediateOrCancel"* execution model and "*Maximum Drawdown Model*" for risk management for our algorithms.

3.7. Algorithm Scoring

Performance measurement is built in QC platform, and is based on the standard fund scoring, including the following:

- **Compound Annual Growth Rate (CAGR):**

$$CAGR = \left(\frac{EB}{BB}\right)^{\frac{1}{n}} - 1\,, \tag{3}$$

where: EB=Ending balance; BB=Beginning balance; n=Number of years (1y=252d).

- **Drawdown:**

A drawdown is a peak-to-trough decline during a specific period for an investment, trading account, or fund. A drawdown is usually quoted as the percentage between the peak and the subsequent trough.

- **Sharpe Ratio:**

$$Sharpe\ Ratio = \frac{R_p - R_f}{\sigma_p}\,, \tag{4}$$

where: R_p = return of portfolio; R_f = risk-free rate; σ_p = st. dev. of the portfolio's excess return.

4. Analysis

4.1 Smart Beta Strategies

Following smart beta results are measured from 1st November 2015 to 1st November 2019 for S&P 500. We have examined the performance of 4 different portfolios (Value, Quality, smart beta with ada boost and smart beta with random forest) on different parameters.

In terms of compounding annual returns, Quality smart beta strategy has performed better than other strategy. Quality smart beta is the winner in terms of drawdown which has recorded the lowest drawdown for the same period. In almost all the parameters which we have observed, Quality Smart beta is the winner and our algorithm is able to identify quality stocks from S&P 500 composites for a given period. We have observed that Ada boost and Random forest are quite close in performance for given criteria so it is difficult to identify winners from these for a given period. We might need to run both algorithms for a longer period. We should also mention that for the same period S&P 500 showed better results with CAGR 12.0%, although with drawdown 12.3%, Sharpe 0.93.

Table 2. Summary Statistics for Smart Beta algorithms

Measurement	Value	Quality	Ada Boost	Random Forest
CAGR	4.2%	8.5%	7.9%	7.4%
Net Profit	17.9%	38.5%	35.8%	33.0%
Sharpe Ratio	0.55	1.05	1.00	0.92
Alpha	0.036	0.069	0.063	0.06
Annual Standard Deviation	0.065	0.066	0.065	0.066
Drawdown	10.5%	9.9%	12.7%	12.0%

4.1.1 Value Smart Beta Strategy

For given formula and approach that we have described above for Value strategy, in terms of risk adjusted returns, Value smart beta seems to be slightly outperformed the benchmark. It has been seen that during Q1 of 2018 Value strategy just followed marked and resulted in loss for market correction and US government shutdown impact. Similar observation for Q42018 where our Value strategy failed to outperform that market. From this it seems that value factor exhibits high sensitivity to the overall macro environment. Value Smart Beta Strategy returns distribution is left skewed with less frequent big positive returns which reflects sensitivity to microeconomic factors.

Figure 3. Value Smart Beta Strategy Equity

4.1.2 Quality Smart Beta Strategy

As mentioned earlier from our overall statistics, Quality Smart Beta strategy is the winner on all our benchmarks. It has CAGR of 8.5% and sharpe ratio is 1.053 which is quite good the underlying reasons why quality stocks tend to outperform the market in risk-adjusted terms are unclear but it argued that cash flow fundamentals steer stock prices more than macroeconomic variables meaning that a well-run firm can gain a competitive advantage through careful capital management. This would in turn minimize the risk of

over-capitalization or over-leveraging, which subsequently affects the stock price positively [26].

Our strategy is able to identify the bad time when macroeconomic conditions start to deteriorate, the more investors will become risk-averse and start investing in less risky and quality stocks and hence quality stocks starts to outperform. Return distribution and drawdown results represent the same understanding.

Figure 4. Quality Smart Beta Strategy Equity

4.1.3 AdaBoost Machine learning based Smart Beta Strategy

We have performed machine learning on the same fundamental data by combining value and quality strategy to identify the signals. We have used Simple Moving Average (20-60 days SMA) as a signal to train the algorithms. In comparison to other strategies, AdaBoost comes at second number in our list of strategies on all the parameters that we have tested.

As we mentioned earlier in Value Smart Beta analysis, for Q12018, Ada boost suffered from same reasons and resulted into same sensitivity to macroeconomic conditions but performed better in FY2017 because of Quality variables from our data which resulted into good returns and able to mitigate the loss of Q12018. We observed the same reason for Q42018 as well. This is also shown by risk adjusted return parameter where sharpe ratio is quite close to 1 and quality strategy sharpe ratio.

We understand that we need to search and add more data and variables to identify and suppress the impact of macroeconomic conditions at the start of the period so that this strategy can learn from value variables and benefits by generating high alpha and sharpe ratio. This can be done as a future scope of project.

Figure 5. AdaBoost (ML based) Smart Beta Strategy Equity

4.1.4 Random Forest (RF) Smart Beta Strategy

As mentioned above, we trained performed Random Forest algorithms on the same fundamental data to enhance the returns by combining both value and quality strategy. We observed that without tuning the Random Forest parameters, it lacks behind the AdaBoost in terms of performance. Random Forest has similar variability as mentioned in statistics but lower returns.

As mentioned in AdaBoost strategy results, Random Forest faced similar issues and required more data and variables to mitigate the impact of value strategy and enhance the returns.

Figure 6. RF (ML based) Smart Beta Strategy Equity

4.2 Random Forest (RF) DollarBars Triple Barrier strategy

Initially, we choose QC platform to tackle S&P 500 tick data. Although, with limited QC processing capabilities it takes ages to collect, not to mention create dollarbars and triple barrier labels. In order to improve processing power as well as to apply ready to use mlfinlab package [27] on laptop, the following research and analysis are based on MOEX blue chips companies, which has a tick data available from one of the russian brokerages. The note we should state, that some of the assets for this research are traded on several stock exchanges (i.e. LSE, VI, ADR, etc.), that definitely add some breaches in data flow

sourced from MOEX. Finally, 16 MOEX constituents which are blue chips and liquid enough were analyzed.

To build a trading strategy, firstly, each asset raw tick data was adjusted to dividends paid and, secondly, processed into dollar bars, which size was determined to have an average daily amount of 10 bars. Thirdly, triple barrier was labeled with set choices of horizontal barrier (1 - 50 days) and minimal returns (0.1% - 3%). Finally, data-series were fed into QC online platform as well as into locally installed LEAN Engine to run a Random Forest based asynchronous trend following trading strategy.

Analyzing triple barrier parameters by comparing F1 score for trade decision, we came to quite expectable conclusion that RF performs better with higher return requirements. Although it leads to a decrease of overall amount of labels, which diminish the investment opportunities as well as leads to ML overfitting (see jupyter notebook analysis on the github).

Algorithms in this part of the research use data series from 1st May of 2009 to 1st December 2019. We have shown here the performance of 2 different portfolios (with vertical barrier 25 and 35 days, which are long only). We found that designed algorithms are better suited for long side investment, and adding a short part is detrimental. Probably, this could be improved by introducing two algorithms, each of which for deciding on long and short trade separately.

We should mention also that both presented algorithms outperforms MOEX with better Sharper, higher annualized return, and lower drawdown.

Table 3. Summary Statistics for DollarBars Triple Barrier RF algorithms

Measurement	Long only, vert. barr. 25 days	Long only, vert. barr. 35 days	MOEX index (benchmark)
CAGR	12.3%	13.5%	11.1%
Net Profit	241%	280%	205%
Sharpe Ratio	0.67	0.73	0.55
Alpha	-	-	-
Annual Standard Deviation	0.162	0.159	0.195
Drawdown	32.9%	29.8%	33.4%

As shown in figure 7, since 2011 to 2015 the strategy showed almost no progress, which can be caused by a limited amount of data to learn algorithms as well as changing economic environment in 2013-2015: high volatility and drop of ruble to dollar exchange rate, high increase of key from 5 to 17 in Dec 2014, and drop of the oil prices, which are highly influential for major oil-producing country. However, since 2015 the presented strategy consistently outperforms the MOEX index.

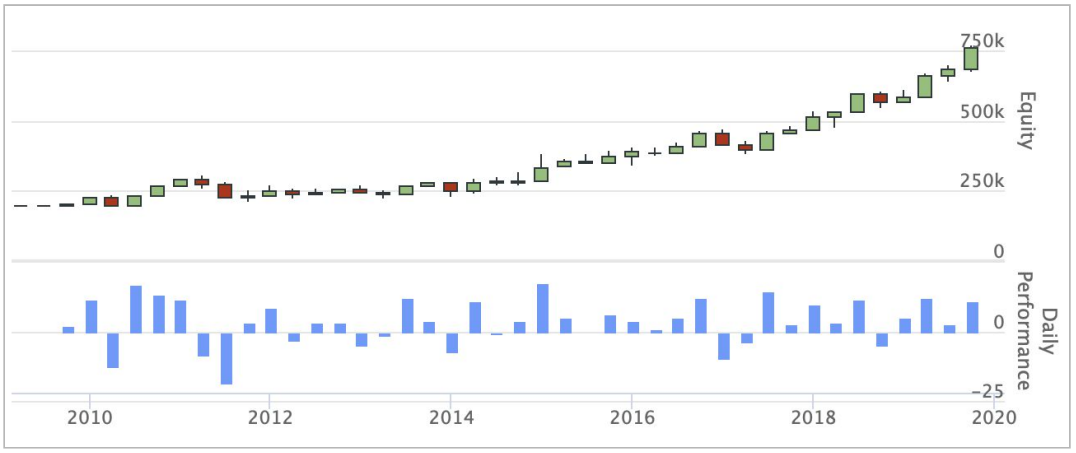

Figure 7. Performance - RF Triple Barrier Strategy

This results show potential for the retail investor to utilize such approach and trading strategy, especially with fee-free opportunities. However, the presented approach is too time-consuming, has many implementational and technical obstacles to overcome (i.e. we met few not precisely documented QC platform functions, as well as a bug in the reporting, which was kindly and timely resolved by the platform support).

5. Concluding Remarks

Based on this empirical analysis that we have performed, 4 out of 5 strategies have performed better in terms of risk adjusted returns. These results suggest that both exposure to smart Beta factors as well as advances in raw tick processing can in fact lead to significant improvements in risk-adjusted returns. When attempting to analyze our results through the lens of academic theory and literature, we find that Smart Beta investing doesn't seem to always abide by the rules. Our historical data don't handle survivorship bias and hence our algorithms. We have not controlled the cost associated for exposure to defaulted firms. Since our paper is comparing performance of algorithms against passive investing in cap-weighted indexes or ETFs we haven't considered cost associated with active investing such as active rebalancing and portfolio reconstruction.

The clear historical outperformance of our 4 out of 5 strategies in two different markets shows that retail investor can leverage algorithmic trading and AI based models using new technologies to generate alpha. Our conclusion is retail trader can outperform the market as a whole with exposure to certain risk factors in medium to long run using quantitative techniques and new technologies. Despite this, we advise retail investors to first evaluate the own risk appetite and technology understanding to have confidence in written algorithms for performance. Also, remember that past performance is not an indicator of future results.

Appendix A

Standardized score (Z-Score)

As this thesis, using data of 500 firms, the large data pool allows us to assume that the values corresponding to individual firms follow a normal distribution, centered around a mean, μ, forming the classic bell-shaped curve below. Normal distribution implies that 68 percent of observations are within one standard deviation, 95 percent within two standard deviations and 99.7 percent are within three standard deviations from the mean.

We assign individual standardized scores z_i, to each underlying variable of interest by subtracting the underlying variable mean from the individual value, x_i. The last step involves dividing the difference between the underlying variables value and the mean value by the standard deviation of all underlying variables. This procedure gives us a standardized number, allowing us to treat all underlying variables similarly, regardless of any discrepancies in units.

$$z_i = \frac{x_i - \mu}{\sigma}$$

All standardized scores for each underlying variable are given equal weights, $\frac{1}{n}$,and added together to compile an overall standardized factor specific score, Z.

$$Z_i = \frac{1}{n}(z_i + z_{i+1} + z_{i+2} + ... + z_n)$$

We now possess a large data set of firms with corresponding overall standardized scores based on selected underlying variables predetermined to reflect a Smart Beta factor. Lastly, the data set is sorted by scores, where the firms with the highest scores are on top of the list. The top 100 firms are selected to proceed to the weighting process, where it will be determined how much weight each firm will have in the final portfolio. The firms with scores below the top-100 are discarded [26].

The Winsor Method

Winsor is an applied statistics method designed to limit the effect of outliers by adjusting their value to an outer limit. Our data set is Winsorized on the overall standardized scores of individual firms, and the limit is set to target scores that exceed three standard deviations from the mean.The weight, i.e. amount invested in each firm is based on their score and its contribution

to the sum of all scores:

$$W_i = \frac{Z_i^{winsor}}{\sum_{i=1}^{n} Z_i^{winsor}}$$

Where w_i denotes the weight for security *i*, and Z_i denotes the factor specific score for security *i*.

Appendix B

Triple Barrier Meta-Labelling Method

Instead of labelling bars from data-series when them reach fixed rate daily (or other period) return, the Triple Barrier Labeling method is applied for this research [3]. Particularly, for each bar the label is assigned when the asset price touches any of three barrier (upper/lower horizontal, or vertical) - see Figure 6. Upper (profit taking), and lower (stop loss) horizontal barriers are scaled by historical volatility of daily returns (particularly, exponential moving average volatility).

For the purpose of this research we applied triple barrier method for making yes-or-no confirmation decision to bet (meta-labelling), when decision on the side was already done by simple moving average method.

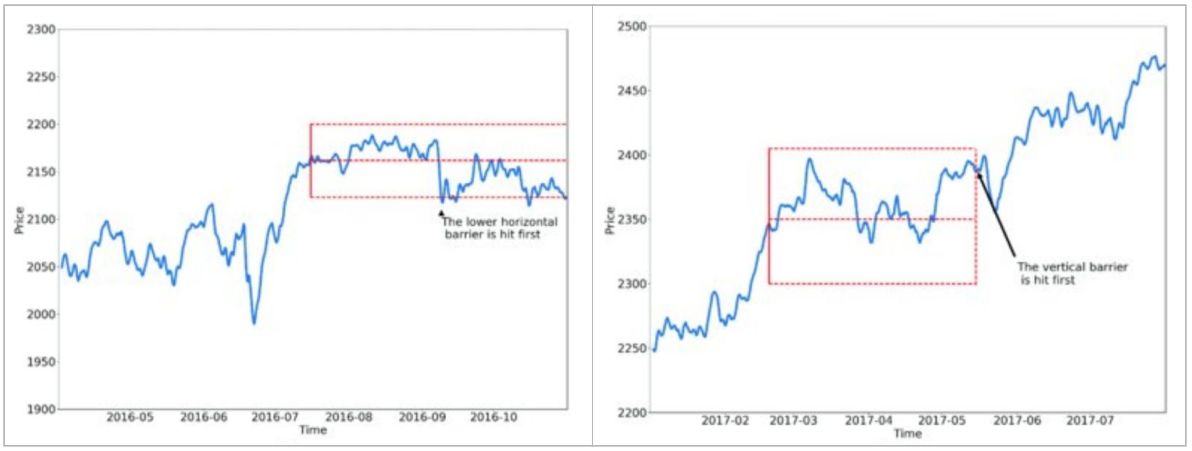

Figure 8. Triple Barrier Method (two alternative configuration). Source - [3]

Appendix C

Implementation of all our above strategies are hosted on GitHub at below path. Also,detailed results of strategies that we have ran as a part of our analysis and conclusion are hosted under documents folder on GitHub.

GitHub Repository Path:

https://github.com/WQU-MScFE-Capstone-MGS/retail-investor-strategies

Acknowledgments

We would like to thank Ritabrata Bhattacharyya for his guidance in this research project. Also, would like to thank WorldQuant University for providing us with this opportunity and the WQU Student Support team for their invaluable support.

References

[1] "March of the machines: The stock market is now run by computers, algorithms and passive managers", The Economist, [Available at: https://www.economist.com/briefing/2019/10/05/the-stockmarket-is-now-run-by-computers-algorithms-and-passive-managers].

[2] Jeff Desjardins, "How Different Generations Think About Investing", Visualcapitalist [Available at: https://www.visualcapitalist.com/how-different-generations-think-about-investing/].

[3] de Prado, Marcos López, "Advances in Financial Machine Learning", John Wiley and Sons Inc.,, New Jersey, (2018).

[4] Malinova, K., Park, A. and Riordan, R. (2012). Do Retail Traders Suffer from High Frequency Traders?. SSRN Electronic Journal.

[5] Isidore, R. and Christie, P. (2019). Model to Predict the Actual Annual Return of the Investor with the Investors' Behavioral Biases as the Independent Variables. The Journal of Private Equity, 22(4), pp.70-82.

[6] Li, W., Rhee, G. and Wang, S. (2017). Differences in herding: Individual vs. institutional investors. Pacific-Basin Finance Journal, 45, pp.174-185.

[7] KUMAR, A. and LEE, C. (2006). Retail Investor Sentiment and Return Comovements. The Journal of Finance, 61(5), pp.2451-2486.

[8] Richards, D. and Willows, G. (2019). Monday mornings: Individual investor trading on days of the week and times within a day. Journal of Behavioral and Experimental Finance, 22, pp.105-115.

[9] Hoffmann, A. and Shefrin, H. (2014). Technical Analysis and Individual Investors. SSRN Electronic Journal.

[10] Choi, J., Kedar-Levy, H. and Yoo, S. (2015). Are individual or institutional investors the agents of bubbles?. Journal of International Money and Finance, 59, pp.1-22.

[11] Koestner, M., Loos, B., Meyer, S. and Hackethal, A. (2017). Do individual investors learn from their mistakes?. Journal of Business Economics, 87(5), pp.669-703.

[12] Keppo, J., Shumway, T. and Weagley, D. (2014). Can Individual Investors Time Bubbles?. SSRN Electronic Journal.

[13] Li, X., Geng, Z., Subrahmanyam, A. and Yu, H. (2017). Do wealthy investors have an informational advantage? Evidence based on account classifications of individual investors. Journal of Empirical Finance, 44, pp.1-18.

[14] Blonski, P. and Blonski, S. (2016). Are individual investors dumb noise traders. Qualitative Research in Financial Markets, 8(1), pp.45-69.

[15] Easley, D., Lopez de Prado, M. and O'Hara, M. (2012). The Volume Clock: Insights into the High Frequency Paradigm. SSRN Electronic Journal.

[16] Reid, P., and Van Der Zwan, M. (2019). An Introduction to Alternative Risk Premia. Morgan Stanley Investment Management.

[17] J.Hsu, V. Kalesnik, F. Li, "An Investor's Guide to Smart Beta Strategies" (http://www.indexinvestor.co.za/index_files/MyFiles/SmartBeta_InvestorsGuide.pdf).

[18] R. Arnott, N. Beck, V. Kalesnik, "Timing 'Smart Beta' Strategies? Of Course! Buy Low, Sell High!" (2017).

[19] Krishnamurthy, R., 2018. Mean Reversion and Beta-Zero Targeting: A Long-Short Equity Trading Strategy. Available at SSRN 3124271.

[20] A. Alford, "Building Confidence in Smart Beta Equity Strategies"

[21] Fama, Eugene F., and Kenneth R. French. 1993. "Common risk factors in the returns on securities and bonds." Journal of financial economics 33.1: 3-56.

[22] QuantConnect. Algorithm Framework Overview. https://www.quantconnect.com/docs/algorithm-framework/overview

[23] López de Prado, Marcos and Foreman, Matthew, A (2014). Mixture of Gaussians Approach to Mathematical Portfolio Oversight: The EF3M Algorithm (June 15, 2013). Quantitative Finance.

[24] Plessen, M. and Bemporad, A. (2018). A Posteriori Multistage Optimal Trading under Transaction Costs and a Diversification Constraint. The Journal of Trading, 13(3), pp.67-83.

[25] Machine Learning for Algorithmic Trading and Trade Schedule Optimization. (2018). The Journal of Trading.

[26] A. Mikaelsson and M. Nilsson (2017) Smart Beta Factor Investing.

[27] Hudson and Thames Quantitative Research. https://hudsonthames.org