Risk on High Frequency Trading Model

Li Chung Yat, Edison
Hong Kong University of Science and Technology, Hong Kong

1. Abstract

High-Frequency Trading has gained its popularity in trading and market-making activities in investment banks and asset management groups because of its ability to execute huge amounts of trade order in a second and its execution speed to find the best deal in the market. To get a better understanding of risks embedded in high-frequency trading, a signal trading based high-frequency trading logistic regressor is built in this project to perform various in-depth analyses on high-frequency trading. The analyses start with explanations of data structure and preprocessing procedures. Then, in the next section, detailed analyses on the selection of machine learning algorithms and inputting features can be found. Followed by a series of performance evaluations with both the prefixed in-sample dataset and 3 days out-of-sample data to determine the model's profitability and usage capability in the Hong Kong equity market. Lastly, based on the limitations and non-realistic assumptions made in the creation of the high-frequency trading model, observable risks and their corresponding solutions are discussed.

2. Introduction

High-Frequency Trading (HFT) is known as one of the algorithmic trading mechanisms. HFT uses strong programming techniques to build up a complex trading model to execute a large number of orders in a short period or to serve the application of certain trading strategies (James, 2019).

Market makers take advantage of HFT mechanisms to perform market-making activities, such as trading on the market's bid-ask spread. Not only market makers will be benefited from HFT, but there also are 2 major advantages brought along to the market, improves the market's liquidity and raise the market's competition.

In this project, a HFT model is built based on signal trading strategy instead of the classic bid-ask spread HFT model, because of limitation on data and complexity of bid-ask spread HFT model, to demonstrate the implementation of HFT in Hong Kong equity market and to serve the purpose for performing risk evaluation on HFT. China Evergrande Group, stock number 3333.HK listed on HKEX is the trading target of this project because of its high volatility and popularity in the Hong Kong stock market. These 2 characteristics play crucial roles in HFT since HFT requires high market participation and volatile prices to make a profit.

3 Data Preprocessing

3.1 Data Exploration

Given that the HFT model is built based on a signal trading strategy, various equity data fields are collected from Bloomberg Terminal, such as equity close price, volume, and moving averages. There is a detailed introduction and explanation of the collected data in Section 4.3, as most of the collected data are curves and indicators in the stock market and used to build trading signals in the model. All data fields starting from 1st January 2019 to 27th November 2019 are collected on a per min basis. Data before 26th April are excluded and used to imitate historical statistical measures and movement. Data originally are prepared based on a 6 months horizon for the sake of data recency, but a 7 months window is applied at the end to compensate for holidays. On every trading day, a total number of 331 periods are included, from 09:30 to 16:00, excluding the lunch break between 12:00 and 13:00.

3.2 Data Preparation

Row Operations

As most of the fields are directly downloaded from Bloomberg, there is a handful of data cleansing work to be done. The only cleansing work is related to missing data on Bloomberg extraction due to the security have no transaction at a certain time. Based on the assumption on data continuity, all missing data fields at time t have the same value as of period t-1, except volume, moving averages and other parameters with moving window.

As the buying signal (dependent variable) is the minority case, a model built appears strong bias in the classification process. This leads to the domination of non-buying signal and the model will become less insightful and useful. Therefore, oversampling dataset balancing is introduced to increase the number of buying decisions and strengthen the model's classifying ability to put a bid.

Column Operations

The main focus of this part will solely be related to signal preparation. Starting from getting a basic understanding of different trading signals, these signals are backtested on historical data (the excluded set of data) to come up with the best parameters setting for signal generations and applied on an actual in-scope dataset. A detailed explanation can be found in Section 4.3.3.

4. Methodology

4.1 Trading

This section will first go through the assumptions made on the HFT model, then the explanation of the selection of trading target and, followed by analyses of the signal trading strategy. Due to limited knowledge of trading strategies and for the sake of simplicity, a fairly simple signal trading strategy is applied to the high-frequency trading model.

4.1.1 Trading Strategy

Based on studies on comparables movement, equity market indicators and other relatable features, the trading strategy is designed to bet on 1-period (minute) forward price to optimize the competitive edges gained from using HFT. In other words, at time t, if price movement from time t to t+1 is positive, the execution signal (dependent variable) at time t will be labeled as 1, which indicates buying, otherwise, a 0 is assigned. The long position is held for 1 period and sold immediately at t+1. If consecutive 1s are found at time t and t+1, it can be treated as a long position at time t and hold until t+2 because the back-to-back short after 1 period canceled out the long position in time t+1 based on the no bid-ask spread pricing assumption which will be further elaborated in the trading assumptions section.

4.1.2 Trading Signal

A trading signal is a trigger, for either buying or selling of a security or other asset, generated by analysis. That analysis can be human-generated using technical indicators, or it can be generated using mathematical algorithms based on market action, possibly in combination with other market factors such as economic indicators (Chen, 2019).

For clarification, the dependent variable mentioned in the trading strategy section is not a trading signal but a model trainer formulated to serve the defined trading strategy. It only acts as the learning target for the HFT model. Although it includes logical interpretation with mathematical calculation, it is defined as the execution signal or buying signal in this project to avoid confusion.

Most of the market or economical indicators are crawled from Bloomberg Terminal. They went through a series of analysis and backtesting process to generate the trading signal with the best parameters' setting. Signals can be categorized into 3 main sectors: namely volume-based, momentum-based, and oversold-based.

The volume-based signal is related to the study on trading volume shocks. By using the volume traded on time t-1 compare to the historical average, the shocking power of new trade is quantified and converted to a trading signal.

Momentum-based signals aim to identify the best buy in opportunity based on the observed inertia on stock price movement. The Moving Average Convergence Divergence is the sole contributor in the momentum-based signal sector.

Oversold market identifiers are used to create a trading signal related to the current market overselling situation. Relative Strength Index and Bollinger band indicator are the most commonly used indicators to reflect oversold and the overbought situation in the equity market, are used for oversold signal generation in this project.

At the end of this section, it is noticeable that only a brief introduction of the trading signals used in this project is made. Still, Section 4.34 will cover a more detailed trading signal analysis, elaboration of logic and parameter testing in signal creation.

4.1.3 Trading Target

Market researches are done on a variety of stocks listed on HKEX, China Evergrande is a few of the stocks that satisfy both criteria for an effective high-frequency trading. 3333.HK appears to be a quite volatile stock with a 1-day per minute volatility of around 12%. This is a critical characteristic to open up the opportunity for trading on small price movements in a short period using the HFT model. China Evergrande's popularity among Hong Kong investors is another important feature that makes it suitable for applying HFT, its popularity provides high liquidity to ensure the signal trading algorithm can execute properly after either long or unwind signal is returned.

4.1.4 Assumptions

There are multiple assumptions made in the HFT model to facilitate the development of trading signals and demonstration of HFT.

The first assumption made is about the stock price movement occurs every minute, neglecting price changes within any minute. Given the fact that every minute could have multiple trading orders, the price movement is likely to occur within a minute. However, due to the limitation on data extraction from Bloomberg Terminal, it is unfortunate that a trade level HFT model cannot be demonstrated but a per min model instead. This also implies that most of the price related independent variables, like the moving average curve, will behave the same. Still, features, like volume, will experience a 1-period lag to fit into reality. (Section 4.3.4 will cover a detailed explanation on the volume indicator) Thus, an assumption on the price movement is necessary to make.

An assumption on stock liquidity is also made. Despite 333.HK's high liquidity, it is not a guaranteed fact. In order to make trading demonstration and model training progress smooth, China Evergrande stock is assumed that can be sold or buy at any trading period without any liquidity concern in the market.

To facilitate trading profit and loss calculation and trading strategy implementation, the model assumes no bid-ask spread exists in the market such that bid price and ask price of the same security is always the same.

4.3 Modelling

You can find an in-depth introduction to algorithm and features selection, followed by general explanation of various trading signals and their rationales behind can be found in this section. Although signal trading is the main focus in this HFT model, study on comparables is also included to capture the industrial trend, market sentiment, and related industry's movement. A number of indicators widely used in the equity market, like Variable Moving Average, the number of orders, are captured in the model as well to serve different purposes. According to individual t-test in the regression model, features with high p-value are insignificant variables, thus, they were filtered out. What is more, features that exist high multicollinearity were also discarded.

4.3.1 Logistic Regression

Logistic regression is a statistical method for analyzing a dataset with the outcome that is measured with a dichotomous variable (Chandrayan, 2018). It is a parametric model to determine log odds of a binary event given by the following equation:

$$\log\left(\frac{prob(p=1)}{1-prob(p=1)}\right) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_i * x_i\ ; for\ i = 1\ to\ n$$

where
β_0 is the intercept of regressor,
β_i are the coefficient of independent variables x_i and,
n as the total number of independent variables in the model,
cutoff used in regressor is 0.5

Considering the existing trading strategy which either gives 1 for positive price movement or 0 for negative or no price movement, logistic regression appears to be quite suitable for the modeling and prediction on the binary outcome scenario. In short, the trading strategy's output provided the best environment to apply a logistic regression algorithm.

Other than that, independent variables, such as comparables' measurement, market indicators, and trading signals, able to fit all logistic regression algorithm assumptions quite well. First, the assumption on no existence of outliers. From the exploratory data analysis, there are only a handful of data distributed at historical mean plus or minus 3 standard deviations. Still, it is understandable that hardly can a per minute market indicators or binary signals, have frequent massive movement. Multicollinearity assumption is validated by the evaluation of variance inflation factor, please visit Section 4.3.6 for elaboration and review variance inflation factor summary (Appendix III) on the HFT model.

From the logistic regression summary (Appendix I), the low Pseudo R square appears to be quite low in which reflecting lots of outcomes are not explainable by the HFT model itself. The claim is non-deniably true but it makes sense to be true. Given the fact that the equity market is full of unexplainable irrationality and white noises, especially on a per-minute trading level, these effects are very significant. For example, even if China Evergrande stock has experienced a 50-period consecutive positive movement, it does not need to keep on the momentum in the 51st period as suggested by the momentum trading signal. These signals are theoretical which means not always applicable in the irrational equity market in reality. However, it is admittable that the model itself is not capable to capture delayed response in the market. Currently, in the model, time t prediction is solely affected by time t measures. A longer scope could have been applied for the evaluation and analyses done on selected features. It is unlikely that randomness in market price movement can be easily interpreted by a few market indicators or peer analysis. What is more, R square in the binary classification of logistic regression could have many computation issues and lead to extreme R square value. Thus, to effectively evaluate the logistic regression HFT model, it is suggested to review its confusion matrice and the related metrics. Lastly. to test the true capability and performance in the market, 3 days of out-of-sample data are prepared to perform simulation.

4.3.2 Comparables

Comparables are introduced in the model to capture the direction and strength of the stock price movement in 4 different kinds of areas. All figures in this section are presented as a percentage change of stock price from period t-1 to t to capture both the magnitude and direction of stock price change from the comparables. Despite the lagging effect, the trend itself, in reality, does appear to have lagging nature on stock price movement as investors are not that responsive to trend, not to mention, HFT is traded every second or minute (in this project).

Industrial Competitors

Agile Group (3383.HK), Country Garden (2007.HK) and China Resources Land (1109.HK) are the selected industrial comparables. China Evergrande group as one of the most mature property developers in mainland China. Studies on its competitors who are also listed in HKEX can, to a certain extent, reflect overall industrial performance and trends in China, as well as Hong Kong investors sentiment on China property market. From the Logistic Regression Summary (Appendix 1), it clearly shows one-period lag percentage change of competitors' stock price has high significance (with 0 p-values) and strong impacts (as the coefficients they have are comparatively high against other sectors' comparables).

Local Property Developer

CK Asset Holdings (1113.HK) and Sun Hung Kai Properties (16.HK) belong to this sector. The reason to include this sector of comparables is to take account for the measurement of the local property market. Even though China Evergrande has limited exposure to the Hong Kong property market as it has only a handful of projects in Hong Kong, it is critical to consider the local market to reflect the overall property industry performance. As of 1113.HK and 16.HK are the 2 largest market capitalization blue chips, they are the best representatives to be chosen for the local property developer sector. Although their percentage price changes sometimes appear to have non-zero p-value, the p-value is still under an acceptable range (<0.05). The null hypothesis in the individual t-test is rejected and they are said to be significant features in the model.

Related Industry Company

Anhui Conch Cement (914.HK) is the only stock in this sector. There are 2 reasons that make 914.HK specials and being chosen as the representative for this sector. First, it is a cement production company. Cement is critical to property or infrastructure development, in other words, the 2 industries have a high correlation. Second, 914.HK is a China-based company. It has a high possibility to be the cement provider for China Evergrande. Their stock price moving magnitude and direction is highly similar at some point because of their project-based business model. It is believed that when China Evergrande announced a new project, both China Evergrande itself and its cement provider will experience a sharp movement of the stock price. Based on the individual t-test, it shows that it has both strong impacts (high coefficient) and significance (0 p-values) in the HFT model.

Market Indicators

Originally, HKEX (388.HK) and Hang Seng Index Futures composite the sector Market Indicators to capture systematic risk and general stock market sentiment. However, due to the high instability of rejecting the null hypothesis in individual t-test in regression analysis, Hang Seng Index Futures is excluded from the model. 388.HK, however, continuously showed its importance in the HFT model with high T statistics.

The regression result validated the hypothesis made on the positive correlation between comparables and China Evergrande stock price. After over 100 trials of data fitting with random training data selection, the regression coefficients have 100% confidence to show that the existence of positive co-movement with high significance across all 4 comparables sectors.

4.3.4 Signals

Different trading signals are backtested and formulated for the regressor to learn the most suitable conditions to apply the trading strategy. All backtesting results on the out-of-sample data are listed in Appendix II.

Volume Signal

It is a trading signal that arises when volume traded at time t-1 is greater than volume's historical mean plus 0.5 SD.

Volume Signal is the indicator to observe the impact on stock price after volume traded at time t-1. Volume is the only collect field that requires a lagging adjustment. It is because volume data collected from Bloomberg Terminal stated as 09:00 symbolize trade volume from 09:00 to 09:01 but price and curves data stated as 09:00 means price and market data at 09:00. Therefore, trading volume at time t-1 is used to perform analysis at time t.

It is believed that the security price is positively affected by a massive deal. Thus, the creation of the VoIume Signal is a comparison of time t-1 trading volume against its historical mean. Using the first 4 months as historical imitation, historical mean is calculated by a continuously expanding moving window as time moves on. However, solely taking historical mean into account yields bad precision as the signal appears in often. Therefore, to further tighten the classification criteria of a significant buy-in, the logic changed to identify buy-in volume greater than mean plus n standard deviation. This n is the parameter setting needed to be backtested with the out-of-sample data in the past.

Based on the backtesting result (Appendix II), a conclusion is made on the fact that more standard deviations are used in the logic, more accurate and precise the volume trading signal is. It also provides a shred of strong evidence which validated the trading signal's hypothesis on huge trading volume will positively impact the stock price at t+1. However, higher accuracy or precision score do not imply that it is a good signal because recall and correlation are poorly low. Low recall reflects that the signal missed many trading opportunities and low correlation suggests there are fewer co-movements between the trading signal and execution signal. It is also important to review to signal's occupation in the dataset, if the chance to flag a signal is too low, the algorithm will be biased. After consideration of both signal's predictive power and advantages that can bring to the model, the signal is raised when trading volume at time t-1 is greater than the historical mean plus 0.5 standard deviations.

Bollinger Bands Oversold Signal

It is a trading signal that arises when the absolute difference between Bollinger Lower Band and current stock price is smaller than 0.05

Bollinger Bands (BOLL) is a set of 3 lines representing the 20-period simple moving average in the middle and 2 adjacent bands formulated by plus or minus 2 standard deviations from the middle band. When the actual price moves closer to the upper band, more overbought the stock is in the market. Similarly, when the actual price moves closer to the lower band, more oversold the stock is in the market. To explain, oversold is a condition where securities are trading lower than the expected price and have a high potential for a price bounce. Thus, oversold market is the desired condition in this HFT based on the defined 1 period forward long-only trading strategy.

Even though this oversold signal does not match the trading strategy well as immediate bounce cannot be spotted in often even after a lower band breach, it is worthy to keep BOLL, the indicator, for further analyses and come up with a more precise trading signal. Regardless, the overselling situation is an appealing market situation to go in. Therefore, the idea to bet around the lower BOLL band is a more comprehensive solution for the signal generation. An absolute difference within a certain threshold between the lower BOLL Band and the actual price is then applied in the BOLL Oversold Signal. Absolute difference not only consider the situation when price breaches through BOLL Lower Band, but it also covers the situation when the Band is nearly breached. This coverage can be highly important because a nearly breach is also implying an overselling market. If signal arises only when breaches happen, these overselling market's opportunity is missed. The assigned threshold also ensures the current market price is in either just breached or just reformed status. Based on the observations made on a variety of securities, the deterrent of a just breached stock is high. This often led to a small bounce back at first in which fits pretty well for our trading strategy. For a just reformed case, the stock usually carries a large momentum to resume its normal state from the breach, it has high inertia to further raise in the next period. Still, the threshold needs to be backtested to find the best-fitted form.

Similar to VI, the signal should be both significant in the modeling part and have high accuracy to predict the outcome. Threshold equals 0.05 appears to have the best balance between the 2.

Relative Strength Index Oversold Signal

Relative Strength Index (RSI) Oversold Signal triggers when RSI drops below 40.

RSI is a momentum-based oscillator, a line frustrates between 0 to 100, to measure the magnitude of current market price changes to reflect market overselling or overbuying situation. Again, due to the short-selling drawback and trading strategy limitation, trading on the overbuying market is not considered. Traditionally, if the RSI measure reaches 30 or below, the security is said to be oversold. An oversold or, in other words, undervalued security is expected to restore a positive trend or a pullback in price. Based on this rationale, a trading signal is developed to bet on low RSI. Although there are numerous trading signals, such as bullish divergence and RSI swings, built based on RSI. Those signals are more suitable for trading in a longer horizon and did not make good use of competitive advantages brought by HFT.

Backtesting is carried to figure out the best RSI level to be used in the RSI oversold signal creation. Based on the result, RSI cutoff at 40 has higher capability to compromise both improvements of prediction on execution signal and being statistically important to the model.

Multicollinearity concerns are raised as both BOLL and RSI are evaluating on the same objective, the oversold market. Still, the concerned multicollinearity effect can be easily assessed by a variance inflation factor, if the measurement suggests that their correlation is too high, a modification will be implemented or, the worst case, to discard one of the signals. However, using 2 signals at the same time to analyze 1 area can also increase the precision to determine the oversold situation. Sometimes, market indicators are misleading in certain circumstances. Therefore, by considering 2 signals generated by 2 different market indicators can greatly reduce the chance of misconception. Simply speaking, if both signals suggest overselling exist, there is higher confidence to classify the stock's undervaluation.

Moving Average Convergence Divergence Signal

This trading signal will arise when Moving Average Convergence Divergence and the corresponding signal line has a difference smaller than -0.05.

Moving Average Convergence Divergence (MACD) is a trend-following momentum indicator that shows the relationship between two moving averages of a security's price (Hayes, 2019). The MACD in practice is the difference between a 26-day exponential moving average (EMA) and a 12-day EMA. To serve a high-frequency setting, 26-period and 12-day period are used instead. The 9-day EMA of MACD is called the signal line for spotting out buying and selling opportunities in theory. It is suggested to sell the security when the MACD crosses below the signal and, in reverse, to buy the security when MACD crosses above the signal line.

The reason to trigger the signal when the difference between the MACD and signal is smaller than -0.05 can be broken down into 2 parts.

First, signal triggers when negative difference occur. The signal encourages the HFT model to take a long position when MACD crossed below the signal. Undoubtably, this act is in fact trading against the theory. However, by constructing the trading signal is the opposite way of the theory turns out is bringing more benefits to the HFT model. The reason behind this is MACD itself has a severe lagging effect. The possibility in a short-term price reversal after a MACD-Signal line cross is high according to previous researches. Thus, the trading signal raised in this project aims to trade on the high possibility of the temporary price reversal in the short term once crossovers happen. Second, threshold 0.05 serves the purpose to limit the model to only trade on new crossovers.

4.3.5 Others

Bollinger Bands' Width

Betting on BOLL width is a well-known trading indicator called The Squeeze. When BOLL come close to each other, this situation highly restricts the movement of market price as moving average (the middle band) is expected to lie between the BOLL. Price volatility will be lower and implies that applying the HFT strategy is likely to have a bad performance. Conversely, high volatility in price is expected when BOLL moves apart with each other. HFT is then in a more favorable situation to use. This theory is validated by the logistic regressor that BOLL has positive co-movement with the HFT-based execution signal.

Variable Moving Average

Variable Moving Average (VMAVG) is an EMA that automatically adjusts the smoothing weight based on the volatility of the stock price. VMAVG is captured in the model to estimate the "true" price of China Evergrande stock because of its ability to smooth out the stock price to reduce noises based on the current market volatility. This feature performs extremely well because a per min HFT model faces extremely high volatility and severely affected by noise in the market. VMAVG's ability to adjust the stock price with the consideration of volatility provides a good approximation of stock price without noise and excess volatility in HFT.

Tick Size

Tick size is the measurement minimum stock price movement. Larger the tick size it is, more volatile the stock price can be in the next period. It is captured in the model to identify the change of volatility from time to time. As suggested by the HFT model, investors should execute the trading strategy when the tick size is large.

Fear Greed

Fear and Greed is the indicator to show market participants' willingness to pay for the stock. Fear symbolized that investors are hesitated to join the market, or simply, to buy the stock. It will decrease the stock price as the demand in the market is low when fear is high. Sellers need to lower the price in the market to attract buyers. The situation reverses when greed dominates the market. As the logistic regression summary (Appendix I) suggested, the HFT model should take a long position in a fear market meaning that when the Fear-Greed indicator falls below 0 (a negative coefficient in the model). To further interpret this indicator, it basically describes the irrationality in the market. "Buy Low Sell High" is a very basic idea in the business world. However, equity market investors seldom take advantage of this basic idea. Investors lose confidence and hesitate to take a long position when the security price is too low. In fact, it is perfect timing as the lowest price suggest a high possibility of undervaluation.

4.3.6 Variance Inflation Factor

Variance Inflation Factor (VIF) is the most common way to detect multicollinearity in regression analysis. Multicollinearity presents when 2 or more independent variables have a high correlation. It could severely affect the regression result with ignorance (Stepanie, 2015).

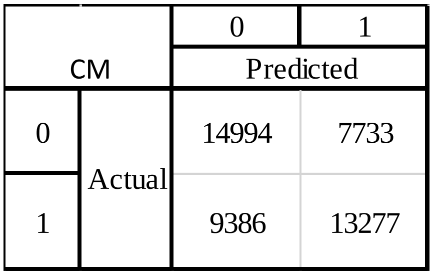
According to Appendix III, although VMAVG appears to have high multicollinearity suggested by the VIF indicator, it brings high significance and strong predictive power to the HFT model. What is more, VMAVG is a moving average, it makes sense that it appears strong correlation with that much price and moving average based signals. Thus, VMAVG is preserved in the model.

4.4 Results

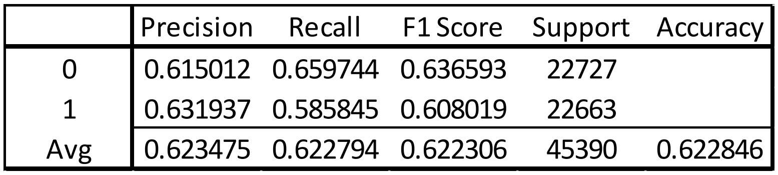
The model achieved a quite satisfactory result, with accuracy, precision and recall achieve over 61% in the balanced training and testing dataset.

4.4.1 Model evaluation

Prediction results on training dataset:



		0	1
CM		Predicted	
0	Actual	14994	7733
1		9386	13277

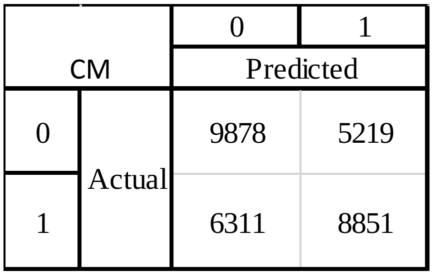


	Precision	Recall	F1 Score	Support	Accuracy
0	0.615012	0.659744	0.636593	22727	
1	0.631937	0.585845	0.608019	22663	
Avg	0.623475	0.622794	0.622306	45390	0.622846

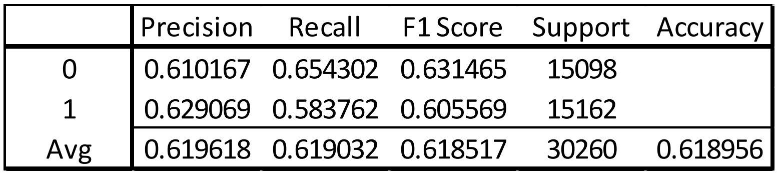
Figure 1. Confusion Matrix on Training Data

Figure 2. Performance Metrices on Training Data

Prediction results on testing dataset:



		0	1
CM		Predicted	
0	Actual	9878	5219
1		6311	8851



	Precision	Recall	F1 Score	Support	Accuracy
0	0.610167	0.654302	0.631465	15098	
1	0.629069	0.583762	0.605569	15162	
Avg	0.619618	0.619032	0.618517	30260	0.618956

Figure 3. Confusion Matrix on Testing Data

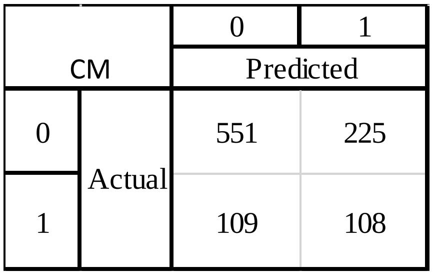
Figure 4. Performance Metrices on Testing Data

4.4.2 Three days out-of-sample market data evaluation

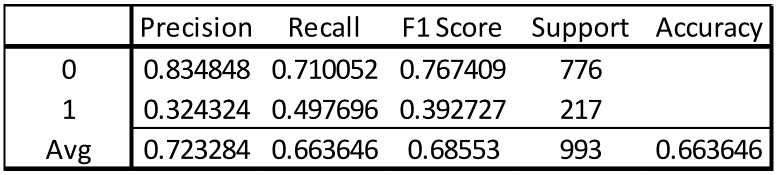
A 3-day out-of-sample data, which is neither included in model training nor historical simulation, from 25th Nov to 27th Nov are used to simulate the HFT model application in the actual market.

For simplicity, the actual trade price is used in the simulation instead of the corresponding bid-ask price. Other than the bid-ask spread, trading fees, stamp duty, and transaction levies are ignored in this simulation. The initial principle is set to be 1 million HKD and only a fixed amount of 2,000 shares are traded (brought) when execution signals arise. A plot of 3-day Profit and Loss and 3333.HK price is attached in Appendix IV.

Prediction results on 3-day out-of-sample market data:



		0	1
CM		Predicted	
0	Actual	551	225
1		109	108



	Precision	Recall	F1 Score	Support	Accuracy
0	0.834848	0.710052	0.767409	776	
1	0.324324	0.497696	0.392727	217	
Avg	0.723284	0.663646	0.68553	993	0.663646

Figure 3. Confusion Matrix on Actual Market Data

Figure 4. Performance Metrices on Actual Market Data

With an overall 66% accuracy and 3-day 10% return, the HFT model itself is a success on a frictionless market. Although return could be much lower or negative in reality, the HFT model proved the significance of the comparables measurement, trading signal and some other market indicators used in the trading model.

5. Risk Analysis

5.1 Systemic Risk

Systemic risk is a risk that a trigger of a certain event would bring unprecedented instability and devastation to the entire industry, or specifically, the entire financial market. The impacts and losses incurred in the 2008 financial tsunami is mainly caused by systemic risk. This risk arises when giants or crucial factors in the industry or financial market fails. When the largest occupier collapses, the whole industry loses its biggest support and crashes as well. Lehman Brothers as an example, Lehman's dominance in subprime and mortgage-backed security business made Lehman occupied the largest percentage of shares in these markets. After its collapse, subprime market after the failure of Lehman was frozen, only extremely creditworthy consumers could get a loan at that time.

Considering the increasing application of HFT in recent years, it bears more and more dependence from the financial market and could very possibly lead to another serious system risk crisis if it falls. According to the International Organization of Securities Commissions Technical Committee, systemic risk is amplified due to HFT activities worldwide strengthen the linkages between financial markets. This means the ripple effect is intensified as the collapse in 1 market is going to bring severer impacts to other countries because of HFT. What is more, volatility is intensified by HFT. When the market is performing extremely unfavorably and disorderly, algorithmic models will perform poorly and push traders to stop using these models. The market could suddenly lose demand and supply from these HFT activities and experience a sharp volatility jump as the bid-ask spread will increase rapidly without HFT. The implementation of HFT also exaggerated the price movement in a short period. Besides, Benchmark trading models will trigger a similar trading signal simultaneously across competitors. This will shock the stock price at a certain moment and further worsen by the existence of HFT. This situation will be even worse when the models are becoming more sophisticated and starting to trade more frequently. The traditional investors will lose their confidence in market activity as they will continuously lose a massive amount of money because of HFT's dramatic shocks on the market.

Systemic risk is embedded in the application of algorithmic HFT. The only way to avoid or mitigate this risk is to reduce the dependence on it or avoid to use it on cross country trading activities.

5.2 Model Risk

Model risk is a risk that occurs when a financial model, which can be a system, quantitative mechanism that relies on either economical or mathematical theories or assumptions to convert data inputs a quantitative outputs, is applied to analyze quantitative information, like market data or financial statements, and the model fails or performs inadequately and leads to unexpected outcomes.

HFT model indeed has considerably high exposure on the model risk because HFT modeling involves quite a lot of coding and assumptions. Using the HFT model in this project as an example, assumptions are made at the initial stage about market data interpretation, modeling as well as trading simulation. These assumptions can be invalid in reality which can lead to a very different result. Trading fiction is the best example that demonstrates the impact it could bring to the model. After consideration of stem duty and trading fees, such a frequently traded model will have a remarkably lower return or even suffer a loss. What is more, as a trading signal based HFT model, signals act as crucial factors in the model. Recalled that the logic and "best" parameter setting behind trading signals are determined by a series of backtesting on historical out-of-sample data. This act, in fact, expose the HFT model to model risk. Despite the fact that these backtesting-generated trading signals yield good performance at this moment, these signals can perform poorly after some time if no adjustments are made from time to time. In other words, the significance of the variables in the model could fade. Either these trading signals generating logic are “outdated” considering market behaviors could be an extreme between various years or algorithm’s assumptions are violated as data changes like 2 low correlation features could exist multicollinearity after few months. Furthermore, the model risk is found in complex coding or data crawling. Given the complexity in designed trading strategies and modeling, Thousands of lines of codes are written and unavoidably might encounter minor coding mistakes. These mistakes could be a typo that is unnoticed until a devastating loss is incurred, or related to the trading model’s design which has low adaptability to new data format. Algorithmic trading also exposes to model risk because of the reliability of external parties’ source or support. The HFT model in this project uses Bloomberg Terminal as a data source. However, if a connection loss happens, the model is not able to trade and hold an unhedged position in the market which might incur a loss as a result. The usage of instruments in the model is another potential area that carries model risk. The model in this project depends heavily on the provided libraries on the Python community, such as sci-kit-learn and pandas. If these libraries malfunction at some point, the model will be shut down and possibly lead to a loss.

To mitigate the model risk, regular reviews and performance evaluations are suggested so as to ensure the model’s suitability on the recent market. Amendments on the trading strategy or the feature selection process should be implemented in case of noticeably bad performance for a consecutive period or failures on individual t-test on independent variables are considerably increased.

5.3 Market Liquidity Risk

Market liquidity risk refers to the shortage of market supply or demand on the security market.

Algorithmic trading model often neglect the fact that the demand and supply uncertainty in the equity market in the model training section. It is understood that this concern is hard to be addressed by indicators in the market. Even the Fear-Greed indicator which shows the market eagerness to engage in the stock, it cannot truly reflect the true liquidity of it. Therefore, based on the assumption of a sufficient demand and supply of the stock at any time, the application of the trading model, in reality, is exposing to serious market liquidity risk. This results in the possibility of missing both buying and selling opportunities. A missed buying opportunity is considered as an opportunity cost while a missed selling opportunity can bring a remarkable loss at critical moments, like in a stop-loss unwinding scenario.

Therefore, to cope with the market liquidity risk in the algorithmic trading model, it is suggested to use frequently traded stocks as the trading target. For instance, the blue-chips or the indexes. These underlying assets seldom have insufficient demand or supply in the market because they have a wide customer base, including both individual and institutional investors, and are commonly used for hedging purposes. The HFT model in this project neither use blue-chips nor indexes as the underlying but the company’s fame and Hong Kong investors’ obsession in the property market minimized the market liquidity risk vulnerability of the HFT model.

6. Reference

Chandrayan, P. (2018, Aug 15). Logistic Regression For Dummies: A Detailed Explanation. Retrieved from https://towardsdatascience.com/logistic-regression-for-dummies-a-detailed-explanation-9597f76edf46

Chen, J. (2019, Oct 10). High-Frequency Trading (HFT). Retrieved from https://www.investopedia.com/terms/h/high-frequency-trading.asp

Chen, J. (2019, May 16). Relative Strength Index – RSI. Retrieved from https://www.investopedia.com/terms/r/rsi.asp

Chen, J. (2019, Mar 21). Trade Signal. Retrieved from https://www.investopedia.com/terms/t/trade-signal.asp

Chen, J. (2019, Jul 19). Systemic Risk. Retrieved from https://www.investopedia.com/terms/s/systemic-risk.asp#:~:targetText=Systemic%20risk%20is%20the%20possibility,%22too%20big%20to%20fail.%22

Hayes, A. (2019, Apr 23). Bollinger Band® Definition. Retrieved from https://www.investopedia.com/terms/b/bollingerbands.asp

Hayes, A. (2019, Jul 1). Moving Average Convergence Divergence – MACD. Retrieved from https://www.investopedia.com/terms/m/macd.asp#:~:targetText=Moving%20Average%20Convergence%20Divergence%20(MACD,from%20the%2012%2Dperiod%20EMA.

Kento, W. (2019, Apr 10). Model Risk Definition. Retrieved from https://www.investopedia.com/terms/m/modelrisk.asp#:~:targetText=Model%20risk%20is%20a%20type,adverse%20outcomes%20for%20the%20firm.

Liberto, D. (2019, Aug 26). Fear and Greed Index. Retrieved from https://www.investopedia.com/terms/f/fear-and-greed-index.asp

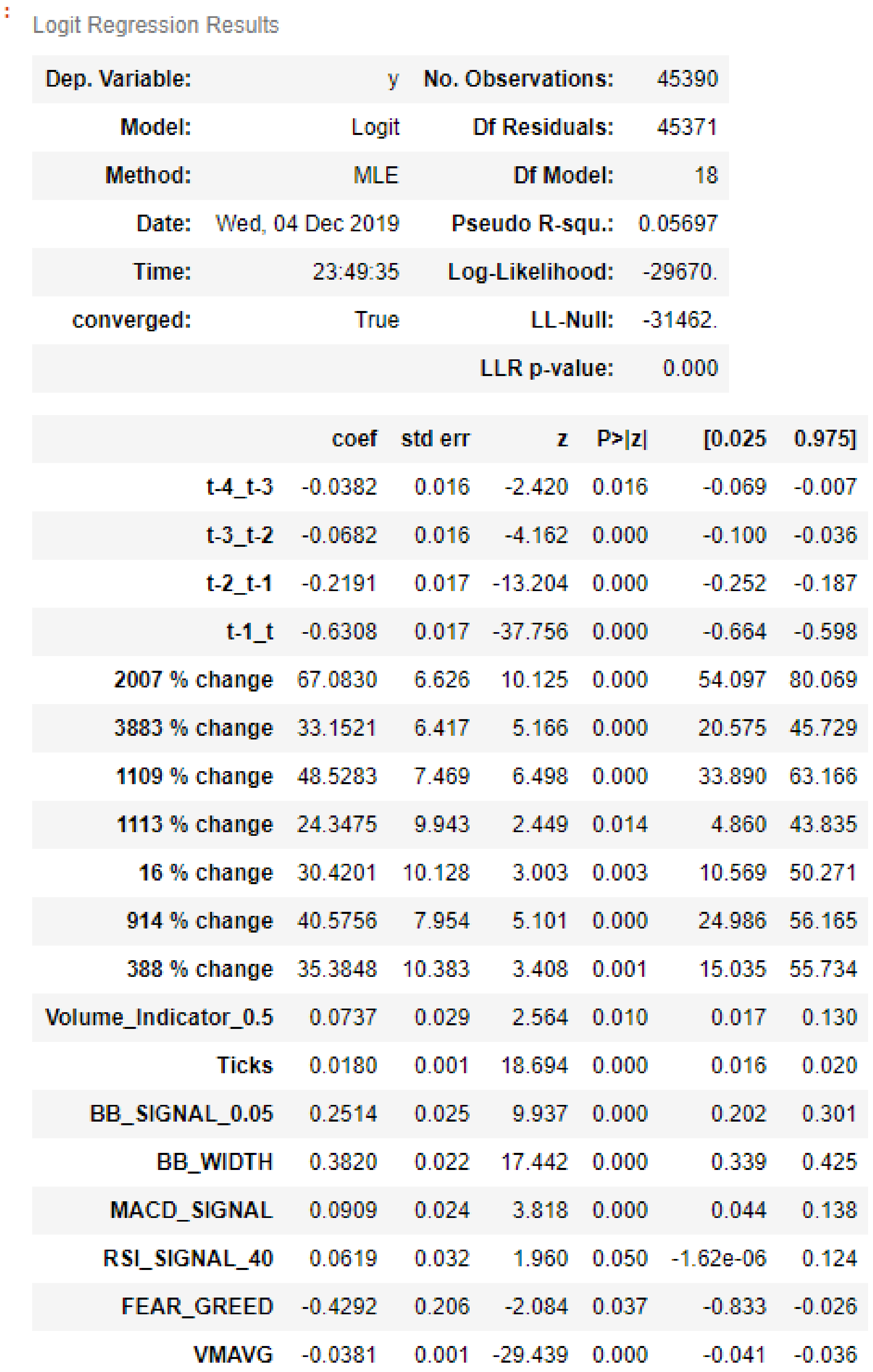
Mitchell, C. (2019, Apr 11). Oversold Definition and Example. Retrieved from https://www.investopedia.com/terms/o/oversold.asp

Picado, E. (2016, Jan 27). Four Big Risks of Algorithmic High-Frequency Trading. Retrieved from https://www.investopedia.com/articles/markets/012716/four-big-risks-algorithmic-highfrequency-trading.asp

StatisticSolutions. What is Logistic Regression?. Retrieved from https://www.statisticssolutions.com/what-is-logistic-regression/

Stepanie (2015, Sept 21). Variance Inflation Factor. Retrieved from https://www.statisticshowto.datasciencecentral.com/variance-inflation-factor/

Appendix I - Logistic Regression Summary

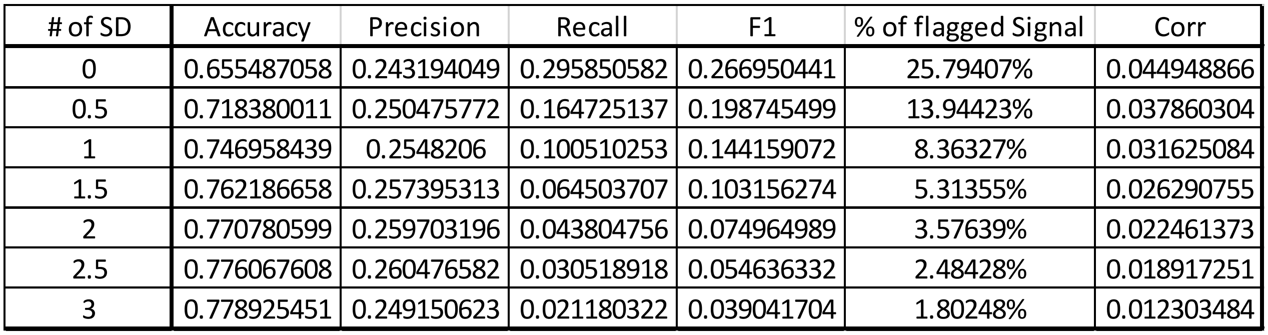


Logit Regression Results

Dep. Variable:	y	**No. Observations:**	45390
Model:	Logit	**Df Residuals:**	45371
Method:	MLE	**Df Model:**	18
Date:	Wed, 04 Dec 2019	**Pseudo R-squ.:**	0.05697
Time:	23:49:35	**Log-Likelihood:**	-29670.
converged:	True	**LL-Null:**	-31462.
		LLR p-value:	0.000

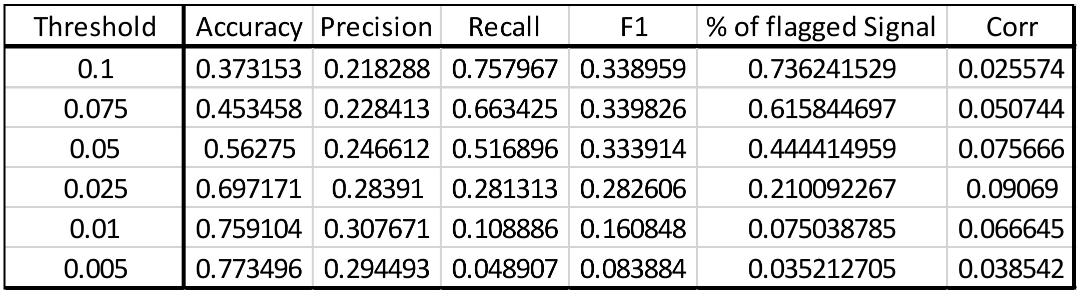
	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
t-4_t-3	-0.0382	0.016	-2.420	0.016	-0.069	-0.007
t-3_t-2	-0.0682	0.016	-4.162	0.000	-0.100	-0.036
t-2_t-1	-0.2191	0.017	-13.204	0.000	-0.252	-0.187
t-1_t	-0.6308	0.017	-37.756	0.000	-0.664	-0.598
2007 % change	67.0830	6.626	10.125	0.000	54.097	80.069
3883 % change	33.1521	6.417	5.166	0.000	20.575	45.729
1109 % change	48.5283	7.469	6.498	0.000	33.890	63.166
1113 % change	24.3475	9.943	2.449	0.014	4.860	43.835
16 % change	30.4201	10.128	3.003	0.003	10.569	50.271
914 % change	40.5756	7.954	5.101	0.000	24.986	56.165
388 % change	35.3848	10.383	3.408	0.001	15.035	55.734
Volume_Indicator_0.5	0.0737	0.029	2.564	0.010	0.017	0.130
Ticks	0.0180	0.001	18.694	0.000	0.016	0.020
BB_SIGNAL_0.05	0.2514	0.025	9.937	0.000	0.202	0.301
BB_WIDTH	0.3820	0.022	17.442	0.000	0.339	0.425
MACD_SIGNAL	0.0909	0.024	3.818	0.000	0.044	0.138
RSI_SIGNAL_40	0.0619	0.032	1.960	0.050	-1.62e-06	0.124
FEAR_GREED	-0.4292	0.206	-2.084	0.037	-0.833	-0.026
VMAVG	-0.0381	0.001	-29.439	0.000	-0.041	-0.036

Appendix II – Signals Backtesting results



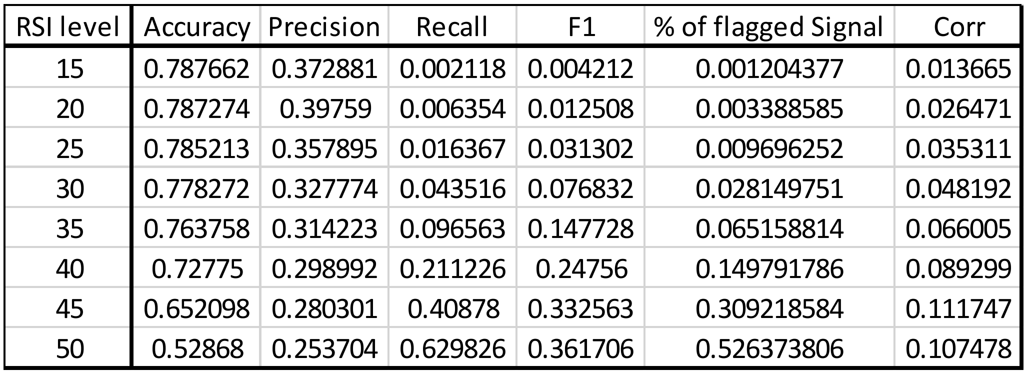
# of SD	Accuracy	Precision	Recall	F1	% of flagged Signal	Corr
0	0.655487058	0.243194049	0.295850582	0.266950441	25.79407%	0.044948866
0.5	0.718380011	0.250475772	0.164725137	0.198745499	13.94423%	0.037860304
1	0.746958439	0.2548206	0.100510253	0.144159072	8.36327%	0.031625084
1.5	0.762186658	0.257395313	0.064503707	0.103156274	5.31355%	0.026290755
2	0.770780599	0.259703196	0.043804756	0.074964989	3.57639%	0.022461373
2.5	0.776067608	0.260476582	0.030518918	0.054636332	2.48428%	0.018917251
3	0.778925451	0.249150623	0.021180322	0.039041704	1.80248%	0.012303484

Figure 1. Volume Indicator Backtesting Result



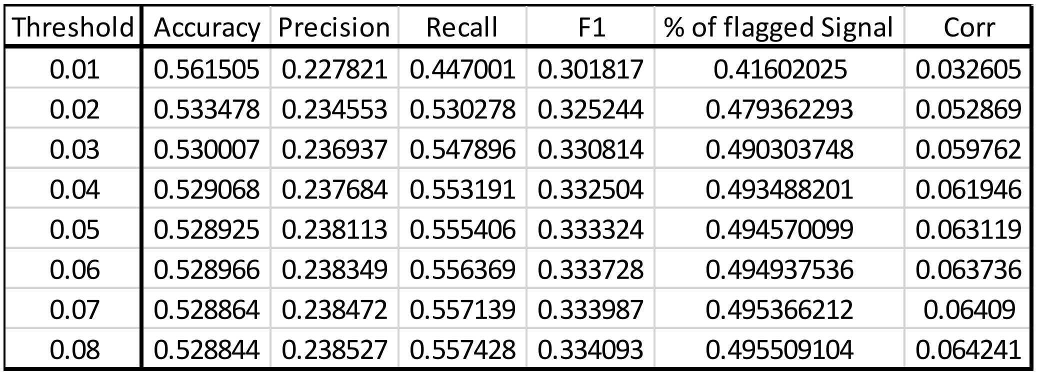
Threshold	Accuracy	Precision	Recall	F1	% of flagged Signal	Corr
0.1	0.373153	0.218288	0.757967	0.338959	0.736241529	0.025574
0.075	0.453458	0.228413	0.663425	0.339826	0.615844697	0.050744
0.05	0.56275	0.246612	0.516896	0.333914	0.444414959	0.075666
0.025	0.697171	0.28391	0.281313	0.282606	0.210092267	0.09069
0.01	0.759104	0.307671	0.108886	0.160848	0.075038785	0.066645
0.005	0.773496	0.294493	0.048907	0.083884	0.035212705	0.038542

Figure 2. BOLL Oversold Signal Backtesting Result



RSI level	Accuracy	Precision	Recall	F1	% of flagged Signal	Corr
15	0.787662	0.372881	0.002118	0.004212	0.001204377	0.013665
20	0.787274	0.39759	0.006354	0.012508	0.003388585	0.026471
25	0.785213	0.357895	0.016367	0.031302	0.009696252	0.035311
30	0.778272	0.327774	0.043516	0.076832	0.028149751	0.048192
35	0.763758	0.314223	0.096563	0.147728	0.065158814	0.066005
40	0.72775	0.298992	0.211226	0.24756	0.149791786	0.089299
45	0.652098	0.280301	0.40878	0.332563	0.309218584	0.111747
50	0.52868	0.253704	0.629826	0.361706	0.526373806	0.107478

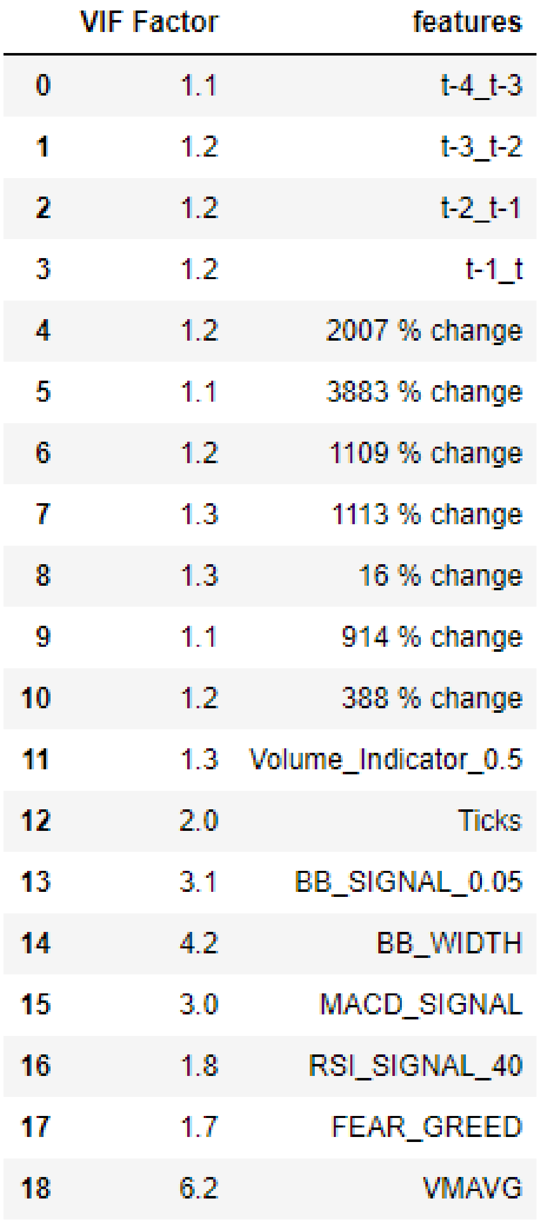
Figure 3. RSI Oversold Signal Backtesting Result



Threshold	Accuracy	Precision	Recall	F1	% of flagged Signal	Corr
0.01	0.561505	0.227821	0.447001	0.301817	0.41602025	0.032605
0.02	0.533478	0.234553	0.530278	0.325244	0.479362293	0.052869
0.03	0.530007	0.236937	0.547896	0.330814	0.490303748	0.059762
0.04	0.529068	0.237684	0.553191	0.332504	0.493488201	0.061946
0.05	0.528925	0.238113	0.555406	0.333324	0.494570099	0.063119
0.06	0.528966	0.238349	0.556369	0.333728	0.494937536	0.063736
0.07	0.528864	0.238472	0.557139	0.333987	0.495366212	0.06409
0.08	0.528844	0.238527	0.557428	0.334093	0.495509104	0.064241

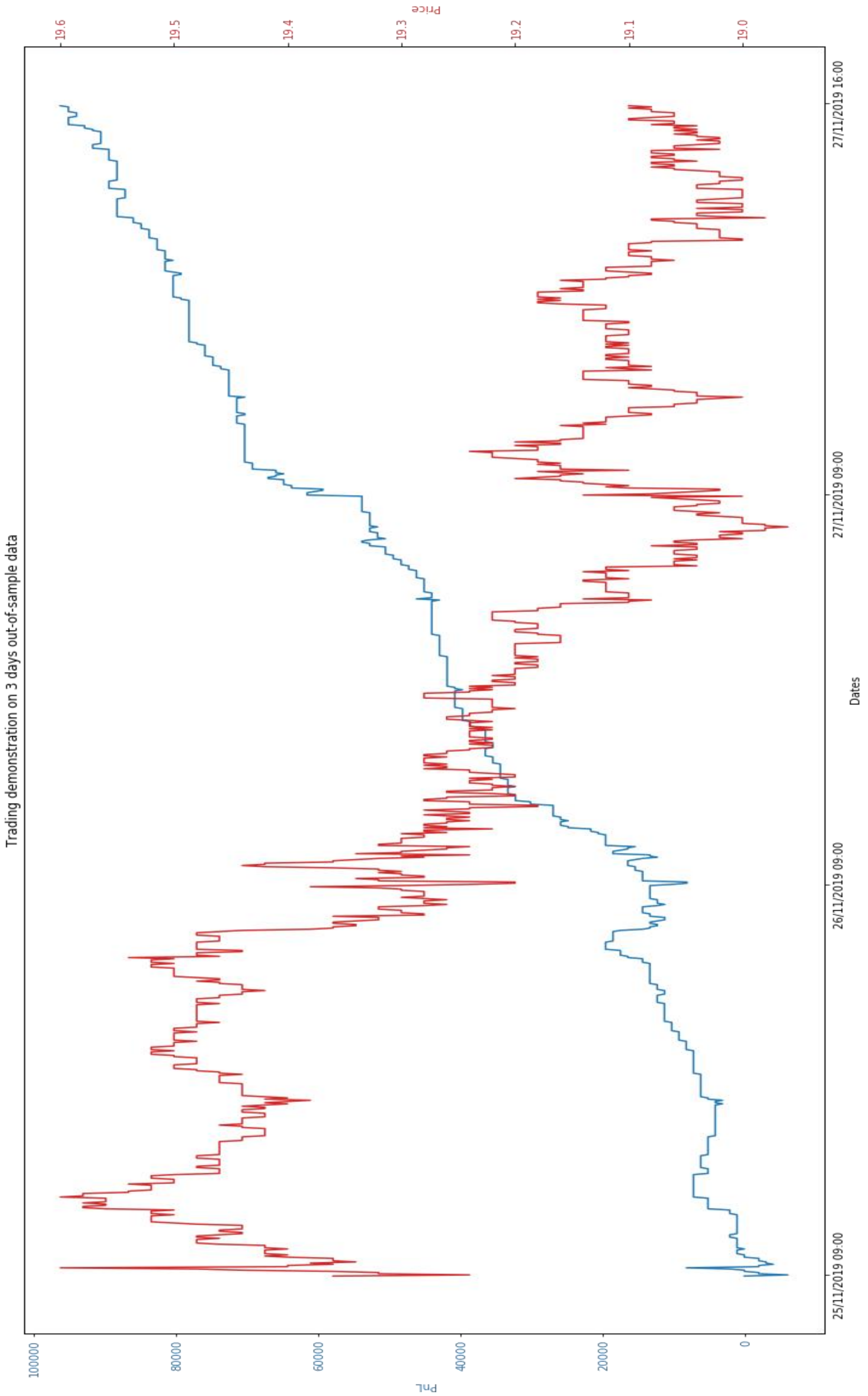
Figure 4. MACD Signal Backtesting Result

Appendix III – Variance Inflation Factor Summary



	VIF Factor	features
0	1.1	t-4_t-3
1	1.2	t-3_t-2
2	1.2	t-2_t-1
3	1.2	t-1_t
4	1.2	2007 % change
5	1.1	3883 % change
6	1.2	1109 % change
7	1.3	1113 % change
8	1.3	16 % change
9	1.1	914 % change
10	1.2	388 % change
11	1.3	Volume_Indicator_0.5
12	2.0	Ticks
13	3.1	BB_SIGNAL_0.05
14	4.2	BB_WIDTH
15	3.0	MACD_SIGNAL
16	1.8	RSI_SIGNAL_40
17	1.7	FEAR_GREED
18	6.2	VMAVG

Appendix IV – Three days out-of-sample market data evaluation


Trading demonstration on 3 days out-of-sample data
Price
PnL
Dates
25/11/2019 09:00
26/11/2019 09:00
27/11/2019 09:00
27/11/2019 16:00