Project Plan

Building a Repository for Teaching Algorithmic Trading

Supervisor: Dr. RuiBang Luo
Woo Chung Yu, Angel
Wu Xue, Snow
Lee Kwan Young

October 9, 2020

Abstract

Thanks to its capacity in managing large data flows and capturing fleeting anomalies in price trends, the use of algorithmic trading has become more prevalent in the financial market. Nonetheless, the development of learning resources for algorithmic trading has failed to keep up with the progression. A large extent of existing resources are predominantly focusing on the US market, and they only cover either end of the tech-finance spectrum. In this project, we propose to build an open-source algorithmic trading code and data repository that puts together the relevant financial concepts and technical skills in implementing an algo trading strategy. The repository shall enable students and investors to form a more comprehensive understanding of what it means to use code to process data, capture statistical patterns and execute trading orders. Leveraging the repository we have built, we will also conduct research on stock price movement prediction with multi-source features, namely exploiting the technical, fundamental, macroeconomic and market sentiment indicators that we created. We will carry out extensive experiments on different machine learning models and algorithms in order to integrate these indicators. The repository is available at https://github.com/awoo424/algotrading.

Contents

List of Figures

List of Tables

1. Introduction

Algorithmic trading[1], is commonly defined as the use of computer programs to automatically make trading decisions, submit orders, and oversee these orders after submission. (Hendershott et al., 2009) Investors are first required to abstract their trading goals into computer instructions or mathematical equations, and then computers would execute the trade orders based on the prescribed instructions.

Advances in technology and telecommunications over the past decade have motivated the digitalisation of financial markets and automation of trading processes. With the introduction of electronic limit order books, algorithmic trading systems have evolved to be an ideal mechanism to organise large flows of information, react to market events readily and to capture statistical patterns within or across financial markets. By now, it is gauged that up to 90% of total trading in the market is algorithmic. (Melin, 2017) Under such a paradigm, it is not surprising to see that big banks, hedge funds and institutional investors have all shifted to using algo trading to execute a trading opportunity in the market. On the other hand, individual investors face a different fate. With a scarcer supply of resources, data, technology and trading experience, algo trading has been accessible but uneasy for individual investors. The large amount of data flow across various sources and the need for timely adjustment of strategies have challenged their cognitive limits and place them at a disadvantage.

1.1 The challenge of learning algo trading

We proceed on to study the learning journey of algo trading, and to understand how it has put a stumbling block in the way of individual investors or students who are intrigued to pick up the skill.

Algorithmic trading is a multidisciplinary field that involves knowledge in three domains, which are finance, programming and statistics. While one could easily look for online learning resources on the Internet nowadays, they still present various challenges to a beginner who would like to learn the field from scratch. As we have illustrated the landscape of existing algo trading learning resources in Figure 1, the problems underlying the present landscape could be summarised as follows:

- There are lots of blog posts and online materials about algo trading, but they are scattered; and are usually just focusing on a few concepts or topics.

- There exist open-source repositories that feature the implementation of a wide selection of algo trading strategies, but they do not come along with explanations and assume that the user knows the financial logic behind.

- On the other hand, there exist websites that provide detailed discussions about financial concepts and rationale of trading strategies (e.g. Investopedia), but they are not accompanied by any code examples.

- Majority of the resources do not fit the local context (e.g. about trade execution in Hong Kong) and are merely focusing on the US market.

[1] Also referred to as “algo”, “robo” or “black-box” trading in the literature

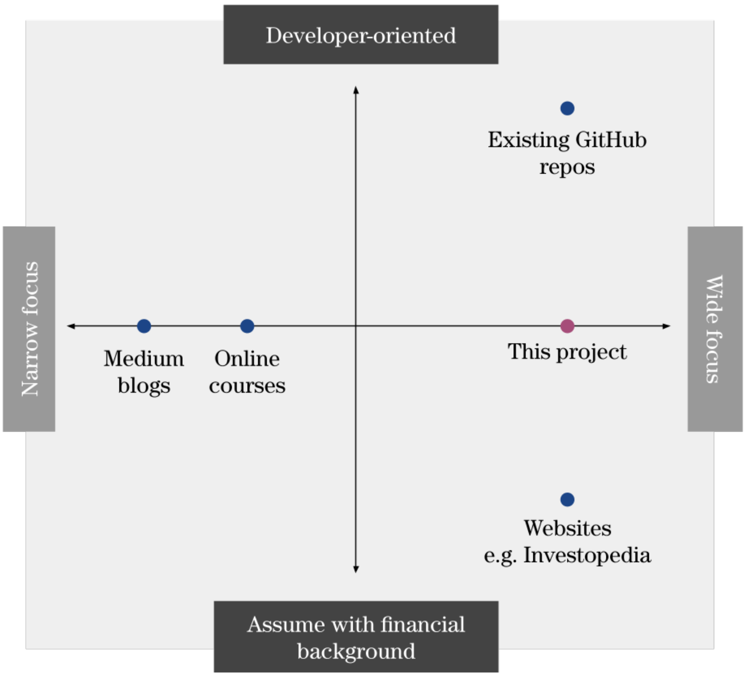

Figure 1: *Perceptual map that illustrates the characteristics of existing learning resources for algorithmic trading.*

As a result, the learning journey of algorithmic trading has inevitably become a scavenger hunt that requires the learner to go back and forth across various websites and online resources. Personally, we also find this process inefficient and daunting. Being a group of students who are majoring in Computer Science and Finance at the university right now, we are eager to dedicate our efforts into creating learning resources that could align with the financial setting of Hong Kong, and could pique the interest of beginners.

1.2 The challenge of applying algo trading

By virtue of this project, we would also like to leverage the resources that we created to conduct research on stock movement predictions. A lot of existing methods have looked into predicting stock trends with single-source features (e.g. purely based on technical analysis). However, in practical algorithmic trading, it is always beneficial to obtain data from different sources, in order to account for different factors causing fluctuations in the stock market. **Stock movement prediction with multi-source features** is a powerful extension in which indicators of various sources (e.g. technical and sentiment) are used simultaneously to analyse and forecast price trends.

Beyond any doubt, technical and fundamental indicators are important metrics that need to be considered. In addition to these conventional indicators, we believe that it is crucial to account for the human factor in predicting market trends. Zhang et al. (2011) and Bollen et al. (2011) have empirically verified that inclusion of public mood dimensions derived from Twitter feeds improves that predictability of stock market indices such as Dow Jones Industrial Average (DJIA) and S&P 500. The incorporation of sentiment indicators generated from Tweets could achieve as high as 82.7% accuracy in predicting the directional movement of stock prices. (Rao et al., 2012)

Besides, we hypothesise that indicators generated based on housing market data also play an important role in predicting the stock market. In fact, the properties and construction sector accounts for over 10% of weighting in the Hang Seng Index (Hang Seng Indexes Company Limited, 2020), and thus could potentially become a source of volatility in the Hong Kong stock market. Nneji et al. (2013) has also empirically proved that there exist three distinct regimes in the housing market - "*steady-state*", "*boom*" and "*crash*", and that the change in regimes is suggestive of fluctuations in macroeconomic variables.

Within the paradigm of prediction with multi-source features, existing papers have looked into the use of Multiple Kernel Learning (MKL) (Deng et al., 2011) and other machine learning models (such as regression analysis) (Wu et al., 2012) to combine technical and market sentiment indicators in order to predict stock prices. However, there is still a lack of research on approaches that are capable of incorporating features from more than two sources to predict market trends. Moreover, only a limited amount of research efforts have been devoted to investigating the application of these indicators in the Hong Kong stock market.

2. Objectives

This project aims to achieve several educational and research objectives in support of teaching algorithmic trading and stock movement prediction based on multi-source features.

2.1 Educational objectives

The primary educational objective of this project is to build an open-source repository that puts together the financial concepts, data science skills and algorithmic design techniques that are crucial to learning algorithmic trading; and thus to present them in a clear, understandable manner. The contents of the repository aim to appeal to a wide audience and especially to students who have no knowledge in finance, but are intrigued to have a better understanding of how financial concepts are translated into computer code.

2.2 Research objectives

The research objective of this project is to experiment and design an algorithm that integrates indicators of different dimensions in order to predict price movement of stocks listed in the Hong Kong and United States (US) markets. These indicators include technical, fundamental, macroeconomic and market sentiment, and they intend to cover the majority of factors affecting the stock market. The correlation between various indicators and stock market trends will be analysed and evaluated. In addition, a consolidated database that consists of historical stock tick, property prices data and social media data will also be constructed to serve for future research purposes at the University of Hong Kong.

3. Project overview

The project is divided into two parts. The first part will look into the market from three perspectives - microeconomic, macroeconomic and sentimental. Within each perspective, historical data will be collected and indicators to analyse the financial market will be created. In the second part, methods to predict stock trends that take features from multiple sources will be explored and investigated. The final product is envisioned to be a self-contained repository which consists of the necessary modules to build, test and evaluate an algorithmic trading strategy.

Part 1. Beginning with the microeconomic aspect, both technical and fundamental approaches in analysing a listed stock will be examined. Making use of a historical stock tick database, technical indicators that scrunitise different properties such as momentum and volatility of a security will be built and evaluated. Concerning the latter, fundamental indicators that interpret a listed company's performance based on its financial statements will be created.

Proceeding on to the macroeconomic aspect, we will look into the Hong Kong real estate market. Hong Kong residential market transactions data will be collected from the websites of different estate agents. The collected data will be analysed over geographies of any kind (e.g., districts) in order to study the interdependence between the real estate market and stock market in Hong Kong.

Regarding the sentimental aspect, market sentiment will be analysed through data collected from different social media platforms and from news articles relevant to respective stock markets. As the popularity of social media platforms vary across countries, text mining technologies will be applied to extract data that has the highest relevance to the local market.

Part 2. Integrating the findings from the above subparts, this part studies and develops the algorithms to predict stock price movement with multi-source features. Our model will take the indicators that we created as inputs. The output will be a signal that indicates whether the price trend is positive, negative or neutral.

Summary of objectives and significance

- An all-in-one pocket guide to algo trading, demystifying financial jargon
- A database with historical stock tick, property prices and social media data for research purposes
- Implementation of indicators that analyse the market from microeconomic, macroeconomic and sentimental perspectives
- Research on stock price movement prediction with multi-source features

4. Methodology

In Figure 2, the overall workflow of the project is depicted. The three subparts in Part 1 will be carried out in parallel, and Part 2 of the project will focus on building upon the indicators created from the previous subparts.

4.1 Part 1 - Microeconomic analysis

Data collection. Since the database of historical price data will be used for evaluating our strategies and further research purposes, it is beneficial to collect as much data as possible from different stock markets. We will scrape historical stock tick data from stock markets including NASDAQ, New York Stock Exchange (NYSE), Hong Kong Stock Exchange (HKEX), as well as that of Japan, Shanghai and Shenzhen. In all 6 listed domains, historical prices of 1-day interval will be collected since the initial public offering (IPO) of each listed company, and price quotes of 1-minute interval will be collected from 3rd June, 2020 till the end of the project period. These data will be collected with the use of yfinance module, Yahoo! Finance and BaoStock APIs. They will be organised in csv format, which is the ideal format for data-processing and data manipulation with Python later on.

In addition to the price quotes, historical financial statements (including profit and loss statements and cash flow statements) of listed companies among the aforementioned stock exchanges will be collected and organised in a database. They will be downloaded from various publicly available sources, depending on the stock market (e.g. IR bank for

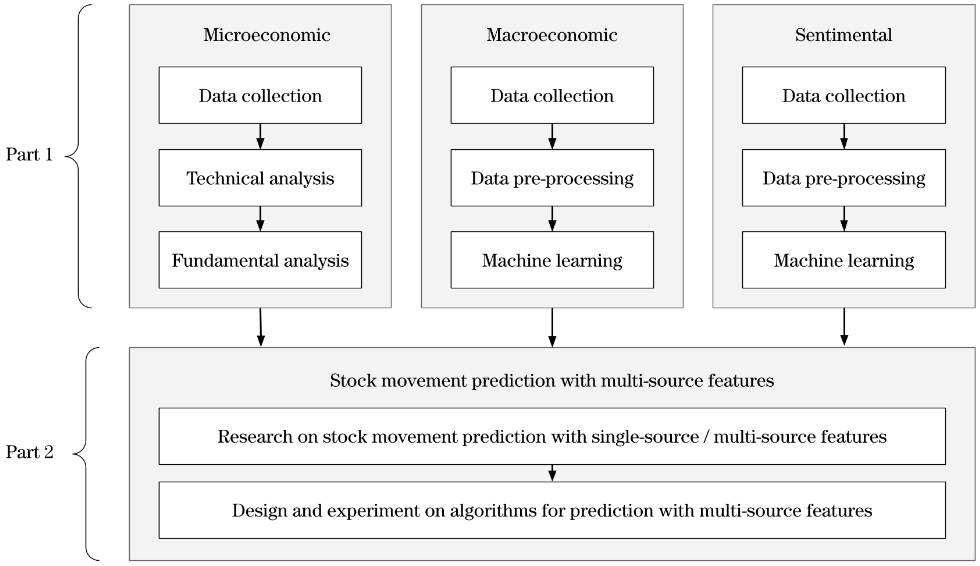
Part 1
Microeconomic
Data collection
Technical analysis
Fundamental analysis
Macroeconomic
Data collection
Data pre-processing
Machine learning
Sentimental
Data collection
Data pre-processing
Machine learning
Part 2
Stock movement prediction with multi-source features
Research on stock movement prediction with single-source / multi-source features
Design and experiment on algorithms for prediction with multi-source features

Figure 2: *Flow chart showing the overall workflow of the project*

Japan). Complementing the APIs, web scraping libraries such as Selenium and Beautiful Soup will also be used to crawl fundamental indicators from Yahoo! finance's website.

Technical & fundamental analysis. We will then use Python (mainly the Pandas and Numpy libraries) to pre-process, clean and analyse the data. Upon completion of data processing, the core technical and fundamental analysis libraries will be coded in two different programming languages at the same time, which are Python and Julia. We choose to create the Julia equivalent of the Python implementations based on the following reasons:

- It has a legible syntax and is easy to learn.
- It incorporates vector notations and DataFrames as part of the language.
- It compiles codes in advance and thus is designed to be fast.
- Despite the lack of community support right now, it is anticipated that it will grow to be an important coding language for quantitative finance.

4.2 Part 1 - Macroeconomic analysis

Data collection. We will collect Hong Kong's current residential property prices and historical residential market transactions data from websites of different estate agents, such as Spacious.hk, Squarefoot.com.hk and Centaline Property Hong Kong. We will either directly download the data from its website, or use web scraping libraries to extract the relevant data. The collected data will then be stored in csv format, and will be cleaned and pre-processed using Python (mainly the Pandas and Numpy libraries).

Machine learning analysis. After pre-processing the data, we will aggregate the price dynamics and carry out feature engineering, for instance by calculating housing price indices over certain regions. Then, machine learning methods will be applied in order to capture price dynamics over different districts or regions of Hong Kong.

4.3 Part 1 - Sentiment analysis

The sentiment analysis part will focus on the US and Hong Kong stock markets, given the following reasons:

- NYSE and NASDAQ are the largest stock exchanges in the world, and a considerable amount of investors in Hong Kong do trade US stocks.
- Hong Kong has recently launched a new NASDAQ-like tech index, and has become more able to attract Chinese companies listed in the US (such as Alibaba (BABA) and JD.com (JD)) to hold secondary listings in Hong Kong.

As both markets are of high liquidity with many active investors, they play an important role in the world financial market. Thus, it is believed that changes in public sentiment across the media are suggestive of stock price movements in these markets.

Data collection. Regarding the US, Tweets and Facebook data will be used to analyse the market sentiment, as they are the most popular social platforms in the country. All stock tickers listed in NYSE and NASDAQ will be monitored. Real-time Tweets will be collected using the Twitter streaming API (Tweepy) throughout the project period. Social media posts and comments (in each post) on Facebook will be collected using the Facebook Graph API. All these data will be stored in csv format.

Concerning Hong Kong, Facebook data and local financial news articles will be collected to examine the market sentiment, and thus to monitor companies listed on HKEX. Unlike the states, Facebook is the sole most popular social media platform in Hong Kong, with a penetration rate of 82% as opposed to 30% in Twitter. (Statista, 2020) Facebook data will be collected using a similar method as for the US. Additionally, we will make use of web scraping libraries such as Beautiful Soup to crawl news reports for different stock tickers. They will be collected through channels including influential newswires[2] and authorised news agencies in the region. These data will also be stored in csv format.

Data pre-processing. As the same with the previous parts, Numpy and Pandas libraries will be used to pre-process and clean the raw data. The sequence of steps in carrying out the pre-processing includes tokenization of text, elimination of stop words, stemming and lemmatisation, as well as building a frequency dictionary. The Natural Language Toolkit (NLTK) suite will be used throughout these procedures.

Machine learning analysis. Different machine learning models, for example as convolutional neural networks (CNN) and BERT (Devlin et al., 2018) will be utilised to classify text into three categories including positive, negative and neutral. The design of training procedure will be referenced from existing research (Sohangir et al., 2018), and we will compare the performance of different models applying on our dataset.

[2] A wire service that electronically provides up-to-date information to the media about breaking news, issues and events.

4.4 Part 2 - Stock movement prediction with multi-source features

Research & experiments. Proceeding on to Part 2, we will first conduct literature review on recent papers related to stock price prediction with the use of indicators from multiple sources. Moving on, we will research on methods to improve upon the baseline performance and run experiments with the historical price data collected.

In parallel with building the code base, we will create a website that features documentation and explanation of relevant financial concepts involved in the code using Sphinx. Table 1 summarises the tools and programming languages that we have selected for each part of the project.

Item	**Tool / Coding language**
Part 1: Microeconomic analysis • Data collection • Technical & fundamental analysis	• Python • Julia • Various APIs, open-source libraries
Part 1: Macroeconomic analysis • Data collection • Data pre-processing • Machine learning analysis	• Python • Various open-source libraries
Part 1: Sentiment analysis • Data collection • Data pre-processing • Machine learning analysis	• Python • Various APIs, open-source libraries
Part 2: Stock movement prediction with multi-source features • Research & experiments	• Python (Pytorch)
Documentation website	Sphinx

Table 1: *Summary of tools and coding language for different parts of the repo*

5. Deliverables

In summary, we anticipate to have the following deliverables by the end of the project period.

1. A **database** with historical stock tick data, housing prices data in Hong Kong, social media and financial news data that reflect market sentiment.

2. An **open-source code and data repository** with the below features:

    - Technical and fundamental indicators for analysing a security
    - Indicators for detecting macroeconomic trends in housing market
    - Indicators for examining market sentiment
    - A pipeline for creating, testing and executing an algo trading strategy
    - Documentation with explanation of code and financial concepts

3. **Survey and research findings** on stock price movement prediction with multi-source features

6. Schedule

The project schedule is detailed in Table 2 on Page 12, and a Gantt chart illustrating the timeline is shown in Figure 3 on Page 13.

Milestones	**Time period**
Phase 1: Ideation and research • Background study on existing resources for learning algo trading • Data collection for all three subparts in Part 1 • Familiarisation with Julia	August - September 2020
Deliverable: Project plan, project website	4th October, 2020
Phase 2: Project development • Part 1 code implementation (50% done) • Literature review • Documentation	October - December 2020
Phase 2: (cont'd) • Part 1 code implementation (100% done) • Part 2 research & experiments	January - February 2021
Deliverable: Interim report, 1st presentation	11 - 24th January, 2021
Final phase: Testing and evaluation • Organise and visualise findings in Part 2 • Final review on code repository • Project exhibition	March - April 2021
Deliverable: Final report, final presentation	19 - 23rd April, 2021

Table 2: *Project milestones*

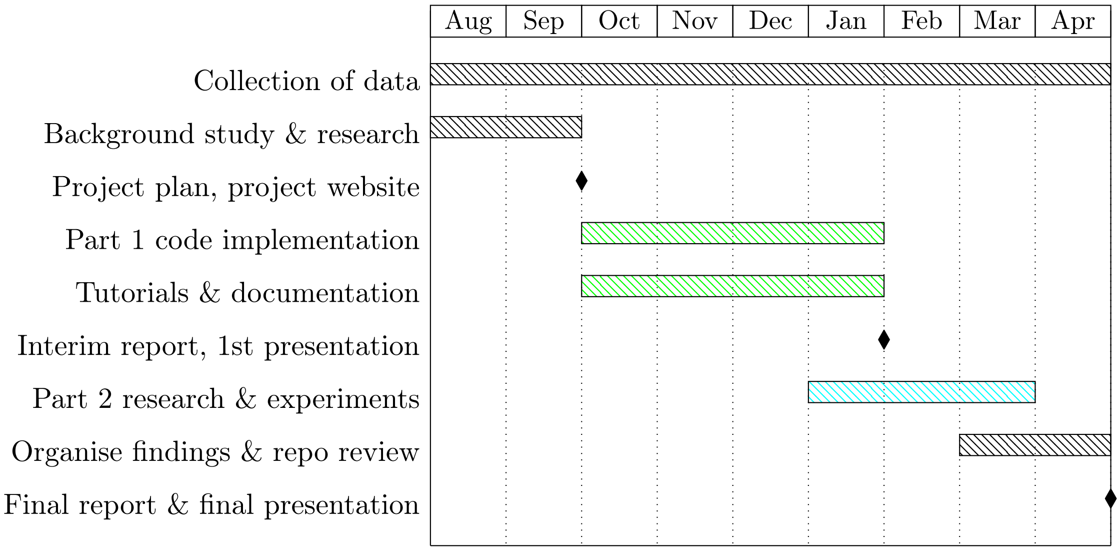

Figure 3: *Gantt chart illustrating the project schedule*

Bibliography

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8.

Deng, S., Mitsubuchi, T., Shioda, K., Shimada, T., and Sakurai, A. (2011). Combining technical analysis with sentiment analysis for stock price prediction. In *2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing*, pages 800–807. IEEE.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hang Seng Indexes Company Limited (2020). Hang seng index factsheet. https://www.hsi.com.hk/static/uploads/contents/en/dl_centre/factsheets/hsie.pdf.

Hendershott, T., Riordan, R., et al. (2009). Algorithmic trading and information. *Manuscript, University of California, Berkeley*.

Melin, M. (2017). Aqr: Computers don't replace human stock pickers, they augment them.

Nneji, O., Brooks, C., and Ward, C. W. (2013). House price dynamics and their reaction to macroeconomic changes. *Economic Modelling*, 32:172–178.

Rao, T., Srivastava, S., et al. (2012). Analyzing stock market movements using twitter sentiment analysis.

Sohangir, S., Wang, D., Pomeranets, A., and Khoshgoftaar, T. M. (2018). Big data: Deep learning for financial sentiment analysis. *Journal of Big Data*, 5(1):3.

Statista (2020). Hong kong social network penetration. https://www.statista.com/statistics/412500/hk-social-network-penetration/.

Wu, J.-L., Su, C.-C., Yu, L.-C., and Chang, P.-C. (2012). Stock price predication using combinational features from sentimental analysis of stock news and technical analysis of trading information. *International Proceedings of Economics Development and Research.*

Zhang, X., Fuehres, H., and Gloor, P. A. (2011). Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia-Social and Behavioral Sciences*, 26:55–62.