Smarter Investment using Big Data, Data Science and Algorithmic Trading

Detailed Project Plan

Chan Chun Hei (3035684908)

Supervisor: Prof SM Yiu

Department of Computer Science

The University of Hong Kong

September 30, 2024

Table of Contents

1. Project Background

Practitioners and academics are continuously developing new and improved techniques to select stocks and increase returns in portfolios. The foundational work by Markowitz (1952) on the Efficient Trading Frontier and Sharpe (1964) on the Capital Asset Pricing Model (CAPM) has laid a solid groundwork for quantitative analysis in financial markets. The development of the Black-Scholes model by Black and Scholes (1973) for option pricing further laid the groundwork for quantitative finance and algorithmic trading.

Algorithmic trading, which involves the execution of orders using automatic pre-programmed trading rules, has been a significant development since its inception in the early 1970s when exchanges began using electronic trading systems rather than manual systems. Early algorithms were relatively simple, executing predefined instructions based on price and volume data. These algorithms laid the foundation for what would later become more sophisticated and intelligent trading strategies. This method has gained substantial traction over the past decades, accounting for approximately 92% of all equity volume in 2019 (Kissell, 2020). Recent studies indicate that the algorithmic trading market was valued at USD 3.1 billion in 2023 and is projected to grow at a rate exceeding 13% from 2024 to 2032 (Global Market Insights, 2024).

Algorithmic trading has significantly influenced financial market dynamics and presents both opportunities and challenge. By executing trades can be executed more efficiently and at better prices. However, it also introduces new challenges by increasing short-term volatility, making the financial market more risky. High frequency trading (HFT), a subset of algorithmic trading, exemplifies these effects by executing a large number of transactions within a short period. This leads to rapid price changes and increase market complexity (Boehmer, Fong, & Wu, 2021).

The rise of data-driven investment strategies can be attributed to advancements in computational resources and artificial intelligence tools. The vast amount of data available on the Internet provides a unique opportunity to enhance investment decision-making processes.

Traditionally, algorithmic trading has relied on quantitative data such as historical prices and volumes. However, with the advancements in natural language processing (NLP) and large language models (LLMs), it has become increasingly feasible to develop trading rules and instructions based on textual data from social media posts and articles.

By combining the analysis on both quantitative and qualitative data, this project aims to develop a novel algorithmic trading strategy that adapts to various market scenarios and outperforms existing algorithms. The report will review the literature on various aspects of algorithmic trading, outline the project's objectives and methodologies, and conclude with the project schedule and milestones.

2. Literature Review

2.1. Types of Algorithmic Trading Strategies

Previous work by Addy et al. (2024) has classified algorithmic trading into various types based on underlying motivations and principles.

One common strategy is trend following, which utilises the momentum of asset prices. This approach often uses moving averages or other technical indicators to identify and follow trends. Zhang et al. (2022) identified two major trend-following strategies, namely the Moving Average Crossover and Volume Weighted Average Price (VWAP).

Another widely used strategy is mean reversion, which is based on the premises that asset prices will revert to their historical mean over time, especially after significant price changes.

Arbitrage strategies exploit price discrepancies between different markets or instruments. The goal is to capture short-term market anomalies, assuming market inefficiencies (Ayala, García-Torres, Noguera, Gómez-Vela, & Divina, 2021). Mean reversion strategies, cointegration analysis, and correlation-based models are common techniques used in statistical arbitrage.

2.2. Nature of Algorithmic Trading

High-frequency trading (HFT) is prevalent in algorithmic trading literature. This phenomenon is explained by Koo (2024). He has shown that traders are increasingly relying on algorithmic advisors for swing trading rather than long-term investing. The impact of HFT on market dynamics, such as price and volume, is also a frequent research topic. For instance, Dutta et al. (2023) discussed how information flow affects the behaviour of high-frequency traders and how certain HFT strategies significantly impact market dynamics including asset prices and transaction volume.

2.3. Machine Learning Approaches

The advent of big data has further enriched the field of algorithmic trading. Extensive financial data, such as historical stock prices, company financial statements, financial news articles, social media sentiments, and macroeconomic indicators are now publicly available online. Machine learning-based algorithmic trading has become a prominent research trend due to its ability to generalise complex patterns and adapt to ever-changing markets. Researchers focus on creating, analysing, and comparing specific algorithmic trading strategies. For instance, N.Dao et al. (2024) examined various deep learning models employed in stock market forecasting, while Majidi et al. (2024) introduced a new approach using reinforcement learning in algorithmic trading.

Large Language Models (LLMs) began gaining popularity around 2017 with the introduction of the transformer model by Google researchers (Vaswani, et al., 2017). Delvin et al. (2018) further built on the transformer model and proposed the BERT model for language understanding. Traditional machine learning models struggle to process and interpret extensive textual data contained in articles and earnings reports effectively. They often overlook nuances that influence market movements. Ni et al. (2024) introduced an approach by employing LLMs to make stock predictions using company earnings reports. In the future, one of the research directions in algorithmic trading and stock predictions is likely to involve LLMs.

3. Project Objectives

The primary objective of this project is to evaluate various algorithmic trading strategies that utilize numerical data. This involves comparing their performance across different markets and investigating their effectiveness in varying market conditions. By identifying the underlying reasons for their performance, the project aims to gain a deeper understanding of these algorithms.

Another key objective is to enhance existing algorithmic trading strategies and explore the effects by incorporating textual data. This involves justifying the principles behind these enhancements and explore the feasibility of using textual data respectively. The project also compares the advantages and disadvantages of the newly proposed strategies to ensure they offer tangible benefits.

Additionally, the project aims to visualise the results using dashboards, allowing us to summarize key insights effectively. By doing so, we can provide a comprehensive view of the collected data and its implications.

Ultimately, the project seeks to develop an integrated investment platform that combines statistical modelling, sentiment analysis and algorithmic trading. By leveraging big data and AI techniques, the platform will provide personalized investment insights and minimize emotional biases, thereby enhancing decision-making processes.

4. Project Deliverables

4.1. Research

The research component of this project will focus on evaluating various algorithmic trading strategies using numerical data across different markets. This includes investigating the effectiveness of these algorithms in different market conditions and identifying the underlying reasons for their performance. Additionally, the project proposes improvements to existing algorithmic trading strategies, justifying the ideas behind these enhancements. The project also explores the feasibility and effects of incorporating textual data into the strategy.

4.2. Software Development

The software development component will deliver a user-friendly application with a graphical user interface (GUI). This application will feature a decision-making dashboard, an algorithmic trading module, and automated article collection and analysis capabilities. Additionally, a database management system will be developed to support the platform.

To ensure the application meets user needs, the project will conduct User Acceptance Testing (UAT) and Usability Testing. Feedback gathered during these tests will be used to refine the application, creating a robust and user-centric investment platform.

5. Project Methodology

The project can be divided into several key components. This section outlines their respective methodologies.

5.1. Data Collection

The project will use historical numeric data (e.g., price, volume) primarily sourced from the Yahoo Finance API. Python libraries such as Pandas and NumPy will be employed for data cleaning, preprocessing, and aggregation.

Textual data, including news articles and social media posts, will be collected via web scraping using BeautifulSoup, Selenium, and Scrapy. Native APIs will also be utilized when available.

5.2. Back Testing

Back testing is a special type of cross validation that involves applying predictive models to historical data to evaluate their viability. This project will use an open-source platform, selected from QuantConnect, QuantRocket, Backtrader, Zipline, PyAlgoTrade, and VectorBT, with a justification for the chosen platform.

5.3. Algorithmic Trading Models

The project will explore the various strategies for predicting asset prices and generating buy/sell signals:

1. Traditional Approaches: Trend-following, mean reversion strategies.
2. Supervised Machine Learning: Logistic regression (LR), random forest (RF), support vector machines (SVM).
3. Reinforcement Learning: Q-learning.
4. Deep Learning: Multi-layer perceptron (MLP), artificial neural network (ANN).
5. Large Language Model (LLM): Generative Pre-trained Transformer (GPT).
6. Multi-Model Approach: Integration of multiple modalities to enhance prediction accuracy.

5.4. Evaluation Metrics

Referencing the work by Cuthbertson et al. (2010) and Sukma et al. (2024), the project will compare and evaluate algorithmic trading models using the following metrics:

1. Annualised Rate of Return (ARR)
   The Annualised Rate of Return is a typical and generally understood metric to measure investment performance. It measures the yearly return of an investment strategy. Different strategies may have different evaluation period. Therefore, the rate of return (R_t) at t is annualised to make it more consistent. A higher annualised rate of return indicates a more profitable strategy.

$$R_{annualised} = (1 + R_t)^{1/n} - 1$$

Equation 1

   Referring to *Equation 1*, the annualised rate of return ($R_{annualised}$) is calculated by annualising the rate of return (R_t) according to the number of years (n) in the evaluation period.

2. Sharpe Ratio
   The Sharpe Ratio is a widely used metric in evaluating investment performance. It evaluates the risk-adjusted return of an investment strategy, helping to determine if returns are due to smart investment decisions or excessive risk. A higher Sharpe Ratio indicates better reward-to-risk ratio.

$$Sharpe\ Ratio = \frac{R - R_f}{\sigma}$$

Equation 2

   In *Equation 2*, the Sharpe Ratio is measured. In the equation, R refers to the return of an investment while R_f refers to the risk-free rate, typically estimated using U.S. Treasury Bond interest rates. σ refer to the standard deviation of the investment return.
3. Win Rate
   The Win Rate is the percentage of profitable trades out of the total number of trades, indicating the competence of a trading strategy disregarding the return of the trades. The Win Rate is calculated using *Equation 3*. A higher win rate suggests a more successful strategy.

$$Win\ Rate = \frac{Number\ of\ Profitable\ Trades}{Total\ Number\ of\ Trades}$$

Equation 3

4. Maximum Drawdown
   The Maximum Drawdown measures the largest loss from a peak to a trough of a before a new peak is achived, as defined in *Equation 4*. It assesses the potential downside risk of a trading strategy. A higher maximum drawdown indicates greater potential loss.

$$Maximum\ Drawdown = \frac{Trough\ Value - Peak\ Vakue}{Peak\ Value}$$

Equation 4

5. Profit Factor
   The Profit Factor measures the ratio of gross profit to gross loss, indicating the profitability of a trading strategy. A profit factor greater than 1 indicates a profitable strategy.

$$Profit\ Factor = \frac{Gross\ Profit}{Gross\ Loss}$$

Equation 5

6. Alpha
   The Alpha (α) measures the excess return of an investment relative to the return of a market benchmark, indicating the value that an algorithmic trading strategy add to or subtracts from the market return. A positive alpha indicates outperformance, while a negative alpha indicates underperformance.

$$\alpha = R - [R_f + \beta(R_m - R_f)]$$

Equation 6

   Alpha (α) measures the unsystematic return of the strategy in *Equation 6*. R is the rate of return of the algorithmic trading strategy, R_f is the risk-free rate, R_m is the return of the market portfolio, and β is its beta. Beta (β) is proportional to the covariance between the strategy return and market return.

5.5. Handling Textual Data

Textual data will be processed using natural language processing (NLP) libraries like NLTK, TextBlob, and Vader. The project will also explore the use of Large Language Models (LLMs) for advanced NLP tasks.

5.6. Integration Strategies

The project will explore various model integration techniques:

1. Hybrid Models: Develop hybrid models that combine traditional and machine learning approaches. For example, use trend-following strategies to generate initial signals and refine them using machine learning models.
2. Ensemble Methods: Create ensemble models that aggregate predictions from multiple models to improve accuracy and robustness.
3. Feature Engineering: Manually or use dimension reduction technique to extract features from raw data, which can then be used by traditional and machine learning models.
4. Real-Time Adaptation: Implement reinforcement learning to adapt strategies in real-time based on market feedback.

5.7. Dashboard

The dashboard will be developed using Plotly and Dash, which are powerful tools for creating interactive web-based visualizations. These tools allow for the integration of various data sources and the creation of dynamic visualizations.

5.8. Testing

Testing will include User Acceptance Testing (UAT) and Usability Testing to ensure the system meets user requirements and is easy to use.

Here are examples of key stakeholders:

1. End Users: Primary users of the app, such as general traders and investors
2. Business Analysts: Professionals who understand the business and industry requirements and can validate if the system meets them. These professionals can be managers from brokers and banks

6. Project Schedule and Milestones

The tentative schedule and phase for the project is shown below in *Table 1*.

Phase/Milestone	Month	Tasks	Deliverables
1 Preparation	Aug 2024	• Brainstorm ideas and confirm topic (10) • Perform background research on algorithmic trading (20) • Conduct market research on investment app (15)	2 Oct 2024 • Detailed project plan • Project web page
2 Planning	Sep 2024	• Consult project supervisor (5) • Define project scope (5) • Define project objectives (5) • Define main features (5) • Research on various algorithmic trading strategies (20) • Prepare the detailed project plan (15) • Prepare the project web page (5)	
3 Implementation	Oct 2024	• Research and implement various algorithmic trading strategies (40) • Perform back testing in different markets and various market conditions (20)	27 Jan 2024 • Preliminary implementation and prototype • Prototype testing • Interim Report
	Nov 2024	• Propose enhancements on existing algorithmic trading strategies (30) • Explore the feasibility of incorporating different textual data (30)	
4 Prototyping	Dec 2024	• Explore the feasibility of related articles collection and summarisation (10) • Design and implement the dashboard to summarize key insights (20) • Design and implement the integrated investment app (20) • Minimal Viable Product (MVP) prototype ready (10)	

5 **Testing**	Jan 2025	• Perform User Acceptance Testing (UAT) to validate the functionality, allowing stakeholders to identify issues or discrepancies and provide feedback for necessary adjustments (20) • Perform Usability Testing to ensures user satisfaction by evaluating the prototype's ease of use involving the user interface (UI) and user experience (UX) (20) • Prepare the interim report (20)	
6 **Fine-tuning**	Feb 2025	• Implement remaining functionalities (30) • Improve the product according to feedback from stakeholders (30)	<u>22 Apr 2025</u> • Implementation of the final product • Final Report
	Mar 2025	• Implement the final algorithmic trading strategy (20) • Cut off back testing and organise the results (20)	
	Apr 2025	• Forward test the final algorithmic trading strategy (20) • Prepare the final report (20) • Prepare for the final presentation (10) • Prepare the poster (5)	

Table 1 Project Schedule

Note: Estimated learning hours for each milestone is indicated inside the parenthesis.

7. References

Addy, W. A., Ajayi-Nifise, A. O., Bello, B. G., Tula, S. T., Odeyemi, O., & Falaiye, T. (2024). Algorithmic Trading and AI: A Review of Strategies and Market Impact. *World Journal of Advanced Engineering Technology and Sciences, 11(1)*, 258–267.

Ayala, J., García-Torres, M., Noguera, J., Gómez-Vela, F., & Divina, F. (2021). *Technical analysis strategy optimization using a machine learning approach in stock market indices*. Retrieved from https://doi.org/10.1016/j.knosys.2021.107119

Black, F., & Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *The Journal of Political Economy, 81(3)*, 637–654.

Boehmer, E., Fong, K., & Wu, J. (2021). Algorithmic Trading and Market Quality: International Evidence. *Journal of financial and quantitative analysis, 2021-12, Vol.56 (8)*, 2659-2688.

Cuthbertson, K., Nitzsche, D., & O'Sullivan, N. (2010). Mutual Fund Performance: Measurement and Evidence. *Financial Markets, Institutions & Instruments, 19(2)*, 95–187.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Retrieved from https://doi.org/10.48550/arxiv.1810.04805

Dutta, C., Karpman, K., Basu, S., & Ravishanker, N. (2023). Review of Statistical Approaches for Modeling High-Frequency Trading Data. *Sankhyā. Series B (2008), 85(Suppl 1*, 1–48.

Global Market Insights. (2024). *Algorithmic Trading Market Size*. Retrieved from https://www.gminsights.com/industry-analysis/algorithmic-trading-market

Kissell, R. (2020). *Algorithmic Trading Methods: Applications Using Advanced Statistics, Optimization, and Machine Learning Techniques.* Elsevier Science & Technology.

Koo, J. (2024). AI is not careful: approach to the stock market and preference for AI advisor. *International Journal of Bank Marketing*. Retrieved from https://doi.org/10.1108/IJBM-10-2023-0568

Majidi, N., Shamsi, M., & Marvasti, F. (2024). Algorithmic trading using continuous action space deep reinforcement learning. *Expert systems with applications, 2024-01, Vol.235*, 121245.

Markowitz, H. (1952). Portfolio Selection. *Journal of Finance*.

N.Dao, H., ChuanYuan, W., Suzuki, A., Sudo, H., Ye, L., & Roy, D. (2024). AI in Stock Market Forecasting: A Bibliometric Analysis. *SHS Web of Conferences, 2024, Vol.194*, 1003.

Ni, H., Meng, S., Chen, X., Zhao, Z., Chen, A., Li, P., . . . Chan, Y. (2024). *Harnessing Earnings Reports for Stock Predictions: A QLoRA-Enhanced LLM Approach*. Retrieved from https://doi.org/10.48550/arxiv.2408.06634

Sharpe, W. (1964). Capital Asset Prices: A Theory of Market Equilibrium. *Journal of Finance, Vol. 19, No. 3*, 425-422.

Sukma, N., & Namahoot, C. S. (2024). Enhancing Trading Strategies: A Multi-indicator Analysis for Profitable Algorithmic Trading. *Computational Economics*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Zhang, L., Wu, T., Lahrichi, S., Salas-Flores, C., & Li, J. (2022). *A Data Science Pipeline for Algorithmic Trading: A Comparative Study of Applications for Finance and Cryptoeconomics*. Retrieved from https://arxiv.org/pdf/2206.14932