THE UNIVERSITY of ADELAIDE

Research Proposal on 'Machine Learning-based Prediction of Stock Market Behaviour'

The University of Adelaide

4336_COMP_SCI_7205A: Artificial Intelligence and Machine Learning

Research Project Part A

Full name:	Pujan Maharjan
Student ID:	a1863495
Email:	pujan.maharjan@student.adelaide.edu.au
Course coordinator:	Dr Alfred Krzywicki
Supervisor:	Haiyao Cao
Submission Date:	11/06/2023

1. Introduction

The study and research about the stock market have been carried out for a very long period and are expected to continue for an extended period. The stock market is a complex phenomenon and is affected by different factors like

- economy of the company, industry, country, or entire world,
- the sentiment of the investors,
- influences of a handful of individuals or organizations.

The anomalies in the stock market - rise and fall in stock prices can be caused by psychological influences and biases that affect the behaviors of investors. The development and advancement of machine learning models have tremendously accelerated the study of the stock market from behavioral perspectives.

Challenge/Research Problem

The key challenge is to identify a single factor that has the highest contribution in the stock price. Most often multiple factors work together that influences the price. Moreover, the data that is used to predict the stock market has a high noise-to-signal ratio. Acquiring quality data that can produce a good signal for stock market prediction is both difficult and expensive. Some market data is available for free, however, most are provided by different vendors with paid subscriptions. In terms of technical challenges regarding machine learning models, the market data (open, high, low, close, volume) is less, and the volume, value, variety, velocity, and veracity of the alternative data like text, images, videos are tremendous to be processed and analyzed by current technologies in real-time. The Reinforcement Learning algorithms have huge potential that can simulate or try to simulate the complex trading environment and learn from the actions taken and rewards obtained.

Research Objectives

The objective of this research is to analyze the influence of the market, fundamental and alternative data sources on the trading strategy that generates signals for trade, optimizes portfolio, and evaluates strategy performance. The majority of focus will be to implement, analyze and evaluate Reinforcement Learning algorithms that are capable of generating alpha. The alpha is defined as the portfolio returns in excess of the benchmark used for evaluation.

Research Questions

The research questions are related to how well the investment model is capable of producing good returns over time.

Q. How does the new investment model perform on profitability? (Yang et. al, 2022, p. 4019)

Q. How does the new investment model perform on reducing risk? (Yang et. al, 2022, p. 4019)

2. Dataset

We plan to use two datasets for the project. The first dataset is the 'optiver realized volatility prediction dataset' which is a 1 second dataset consisting of order data and trade data. This is useful to determine the price, volume and time of the trade. The second dataset is the StockNet dataset consisting of historical price and twitter data. This is useful to analyze the impact of text data on the stock price.

Dataset 1: Optiver realized volatility prediction dataset

This dataset includes order book snapshots and executed trades with one second interval. The data is stored in parquet format specified as book_[train/test].parquet and trade_[train/test].parquet. The order book consists of buy and sell orders of the market. The trade data consists of actually executed trades (kaggle.com, n.d.). The dataset is made available by the Optiver, a proprietary trading firm and market maker, in the kaggle website at https://www.kaggle.com/competitions/optiver-realized-volatility-prediction/data

Dataset 2: StockNet

The StockNet dataset consists of two-year price movements from 01/01/2014 to 01/01/2016 of 88 stocks, coming from all 8 stocks in the Conglomerates sector and the top 10 stocks in capital size in each of the 8 sectors (Xu, 2023). The StockNet dataset was prepared by Yumo Xu and Shay B. Cohen for their paper titled, 'Stock Movement Prediction from Tweets and Historical Prices', 2018. The StockNet dataset is available to be downloaded at their github url, https://github.com/yumoxu/stocknet-dataset.

3. Task Selection and Benefit

Alpha factors are specifically crafted to analyze data in order to anticipate returns within a specified investment realm during a trading timeframe (Jansen, 2020, p. 13). The process of generating alpha includes two phases - research and execution as shown in the figure 1 below.

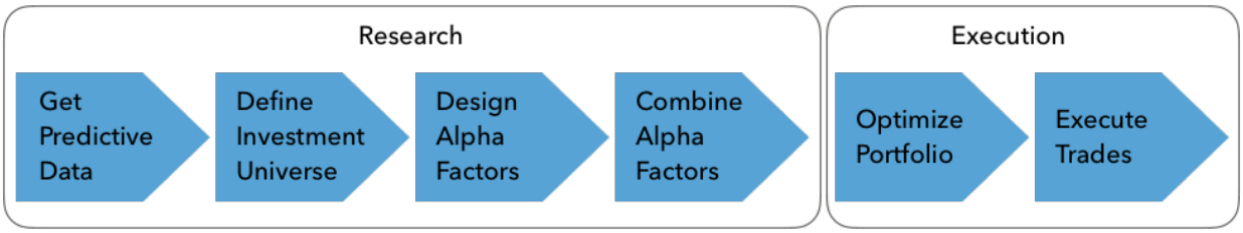

Fig 1: The alpha factor research process (Jansen, 2020, p. 14)

Research: The research phase includes acquiring data containing alpha signals that do not decay too quickly. The investment universe is defined preferably based on theories about financial markets and investor behavior, because many investors still prefer factors that align with such theories. The alpha factor generated from one source of data may be weak, but when combined with multiple sources, it could have a strong effect. The alpha signals include either buy, sell or hold the asset which is then utilized in the execution phase.

Execution: In the execution phase, the portfolio is optimized considering various risk factors and benchmark of evaluation within the scope of the investment objectives. The trading strategies that governs the portfolio should be evaluated which includes backtesting against the historical data and forward testing to validate against new data.

4. Proposed methodology

The workflow for the project will follow a systematic approach to tackling a machine learning project (Jansen, 2020, p. 153). The steps are:

a. **Defining the problems and identifying metrics to measure success** (Jansen, 2020, p. 153)

   We will be using statistical inference to estimate the association of the returns of the asset with a risk factor. The causal inference aims to discover connections in which specific input values result in specific outputs. These inferences are to be used in generating trade

signals and optimizing portfolios. The success of the portfolio is determined by the returns and risks of the portfolio.

b. Collecting, cleaning, and validating the source data (Jansen, 2020, p. 153)

The data is to be downloaded, cleaned, and stored in an appropriate format that will be easy to explore. The text data should be properly assigned timestamps to make it a time series data and align it with the historical data. The data will be stored in HDF or parquet format for quick access.

c. Understanding data by exploring and extracting features of the data (Jansen, 2020, p. 153)

The visualizations of the data using scatter plots show the relationships between the outcomes and the features. This gives insight into the nature of the data and helps in identifying the appropriate machine-learning algorithms particularly suitable for the data. Under feature engineering, we shall apply domain knowledge to transform the data that is most suitable for the algorithms.

d. Selecting one or more machine learning algorithms (Jansen, 2020, p. 153)

The number of assumptions required by the model depends on the type of model. In the reinforcement learning system, the model learns by the hit and trial of the action of the agent. The action is evaluated on the state of the model in a specific environment. In each action, the model receives a reward - either positive or negative. The overall reward determines the performance of the model. In this research project, we shall explore the RL algorithm and compare it with other models.

e. Cross-validate models and hyperparameter tuning (Jansen, 2020, p. 153)

The models selected in the above step are to be cross-validated with different combinations of hyperparameters. The ablation study will be carried out, evaluated, and compared based on error metrics.

f. **Deploy the model and get feedback on its performance** (Jansen, 2020, p. 153)

We shall deploy the model and test it against the demo accounts of different trading platforms. The demo data should simulate the actual trading. The feedback received from this will be utilized to update the model.

5. Ethics statement

The data available in the StockNet dataset does not contain any personal details of the tweeters. The data is anonymized. Thus we can say that the dataset does not pose ethical concerns regarding privacy violations or security. The historical market data contains open, high, low, close, and volume of stock traded on that day. This does not contain any of the crucial information of any party. Thus, the historical data is not subject to ethical concerns.

However, ethical concerns are raised when the alternate data belongs to a user and is used without any consent or when the data is obtained from insider traders. Both of these situations are not applicable to this research project since we are not working with data that has personal information.

6. Expected Outcome and Timeline

The Expected outcome or milestones with timeline for Research Project A are tabulated below:

Table 1: Expected outcome (milestones) for Research project A

Week	Action
Week 5	Develop a literature review
Week 6	Analyze the data and compute results for the baseline
Week 7-9	cross-validate, and tune RL algorithms to achieve SOTA performance
Week 10	Prepare for the presentation
Week 11	Prepare progress report
Week 12	Reflection on the ‘Research project part A’ and planning for ‘Research project part B’

The Expected outcome or milestones and timeline for Research Project B are tabulated below:

Table 2: Expected outcome (milestones) for Research project B

Week	**Action**
Week 1	Revise the planning made in Week 12 of 'Research Project A'
Week 2	Extend the project by synthetically generating more data using TimeGAN or adding more alternate data sources
Week 3-6	Refine the RL models, cross-validate, and tune RL algorithms to achieve SOTA performance
Week 7-8	Develop a research presentation
Week 8-11	Develop the research report
Week 12	Reflection and next steps of the future of the research

References

Jansen, S 2020, Machine Learning for Algorithmic Trading: Predictive models to extract signals from market and alternative data for systematic trading strategies with Python, 2nd Edition, Packt Publishing, Birmingham.

kaggle.com. (n.d.). *Optiver Realized Volatility Prediction*. [online] Available at: https://www.kaggle.com/competitions/optiver-realized-volatility-prediction/data [Accessed 11 Jun. 2023].

Xu, Y. (2023). *stocknet-dataset*. [online] GitHub. Available at: https://github.com/yumoxu/stocknet-dataset [Accessed 11 Jun. 2023].

Yang, M., Zheng, X., Liang, Q., Han, B. and Zhu, M. (2022). A Smart Trader for Portfolio Management based on Normalizing Flows. In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*. Pp.4014–4021.