**SmartTrader**

ISS SLS PROJECT REPORT

ONN WEI CHENG A0092201X

YANG JIESHEN A0003901Y

NIRAV JANAK PARIKH A0213573J

ISA-PM-SLS-2021-01-09-IS02PT

Image from: https://static.vecteezy.com/system/resources/previews/002/062/633/original/business-candle-stick-graph-chart-of-stock-market-investment-trading-on-white-background-design-bullish-point-trend-of-graph-illustration-vector.jpg

**TABLE OF CONTENTS**

[1. Problem Statement 2](#_Toc72072460)

[1.1. PRoblem description 2](#_Toc72072461)

[1.2. approach to solution 2](#_Toc72072462)

[2. model overview and implementation 3](#_Toc72072463)

[2.1. state 3](#_Toc72072464)

[2.2. reward 3](#_Toc72072465)

[2.3. actions 3](#_Toc72072466)

[3. model evaluation 4](#_Toc72072467)

[3.1. features used for evaluation 4](#_Toc72072468)

[3.2. phase 1 4](#_Toc72072469)

[3.3. phase 2 4](#_Toc72072470)

[3.4. phase 3 4](#_Toc72072471)

[4. conclusion 5](#_Toc72072472)

**APPENDICES**

APPENDIX I:

1. Problem Statement

* 1. PRoblem description

A large volume of financial instruments are traded daily on exchanges and a growing majority of these are traded using automated trading algorithms. An automated trading strategy that maximizes profit is highly desirable for mutual funds and hedge funds.

These trading algos can be broadly categorized into:

1. Passive / Fundamental Trading which is primarily a long-term strategy based on market and security fundamentals or passively benchmark index trading.
2. Active / Price Action Trading which is based on buying and selling securities based on short-term movements to profit from the price movements and relies on historical prices (open, high, low, and close) to make trading decisions.

This project will focus on the Price Action Trading which is driven on the characteristics of a security’s price movements. Since it ignores the more subjective fundamental factors and focuses solely on recent and past price data, the Price Action Trading Strategy is more conducive to Machine Learning.

There are essentially two contrasting approaches to developing price trading strategies:

* 1. Model based: This approach attempts to create a mathematical model of the market thru representative variables (eg mean price, corelations) to create a simplified representation of the true complex market model. Some examples of this are trend-following, mean reversion, arbitrage strategies etc
  2. Model free: Here there is no attempt to model the market, rather it looks at price patterns and attempt to fit an algorithm to it. There is no attempt to produce a causal analysis or explanation – just an attempt to identify patterns that will repeat in the future. Examples here are Technical analysis, charting, candle patterns etc

This project will focus on the Model free approach and leverage **Reinforcement Learning** (RL) to develop an automated trading agent (**Smart Trader**) driven solely on historical data and derived technical features.

* 1. approach to solution

While RL has traditionally proved its capabilities in learning to play games, RL presents a unique opportunity to model the complexities of the trading strategy as a game in which an agent can be trained to maximize its total reward from the market.

Trading is a continuous strategy in which decisions are taken sequentially and which collectively culminate into a profit or loss. As such, there are no specific label associated with any historical data and this is a major problem for traditional Supervised Learning (SL) approaches. On the other hand, RL can use feedback from its own action and experiences ie the rewards returned by the stock market to learn an optimal trading strategy.

RL also has different learning goals from SL. While SL learns to make the best predictions (Classification or Regression), RL learns a policy for actions that would maximize its long-term cumulative reward, which is goal of trading.

Trading is a continuous activity and has no defined endpoint. Also trading is a partially observable environment as we do not have complete information about the traders in the market. Since we have a Partially Observable Markov Decision Processes (POMDP) and we don’t know the reward function and transition probability, we use model-free reinforcement learning which is **Q-Learning**.

Q-learning is a model-free reinforcement learning algorithm. It is a value-based method that supplies information (exploit) to an agent for the next trading action (buy, hold, sell). It is an off-policy algorithm as the q-learning function learns from occasional random actions (explore) that are outside the current policy.

**Deep Q-Learning** approximates the Q-value with a neural network as a function approximator.

Q-value refers to the action quality and is stored in Q-table with dimensions [state, action]. The Q-value is the maximum discounted future reward when an action a is performed in a state s. The objective of Q-learning is to learn the optimal value of the Q-table that will provide the maximum reward at the end of the n number of training cycles or iterations. It functions well without the reward functions and state transition probabilities.

At run-time, the agent will evaluate rate each and every action from the current state and select the one with the maximum Q-value

1. model overview and implementation

The analysis and work done in this project aims to improve on an existing Deep Q-Network reinforcement learning model which is available on Kaggle: <https://www.kaggle.com/itoeiji/deep-reinforcement-learning-on-stock-data/notebook>

The model divides the data set into the training dataset and the test dataset. The agent learns using the training dataset and is used to predict buy/sell/hold actions to maximize profits.

A deep neural network is used to predict the action value based on input states. Due to the nature of stock markets, prices are somewhat correlated sequentially. Hence, an experience relay is deployed to stabilize the network. The agent uses a memory to store batches of historical data which is parsed in randomly to train the neural network. The target Q network is updated periodically to further reduce correlation.

The hyper-perimeters are listed in Table xx, and the structure of the neural network is displayed in Table xx

* 1. state

A **state** is the description on the current environment that the agent finds itself. The details on this state are provided to the agent by the environment. It is important that this state has the [Markov Property](https://en.wikipedia.org/wiki/Markov_property) ie the current state information has sufficient and necessary information for the agent to accurately predict expected next rewards and next states given an action, without the need for any additional information.

In our trading model, our OHLC daily price data is a time series and is not stationary ie there is a trend embedded within the time-series. We make the daily OHLC price data by calculating the daily returns ie today’s price – yesterday’s price. This makes the returns data stationary and this is what we use in our model to describe current state.

Furthermore, given the inherent trends in the price data, a single day of price returns is not sufficient to describe the complete state of the market. Hence we use a sliding window of n days of historical returns data (defined by parameter historical\_n = 90) in the state description to provide the agent the full view of market movements in the recent past.

Thus the full state returned by the environment on every step is the today’s return + returns for the historical\_n days. This state is input into the NN network to model the Q-value for the state-action value.

* 1. ACTIONS

Actions is the set of all possible moves the agent can make from a given state. For our trading agent, there are only three permissible action: BUY, HOLD or SELL.

In order to affect the off-policy learning the model employs an Epsilon Greedy Policy to find the balance between the model’s Explore vs Exploit behaviour. The model defines a starting epsilon value (epsilon = 0.9) which starts decreasing by a pre-defined value (epsilon\_decrease = 1e-3) after a predefined number of step (start\_reduce\_epsilon = 200) till it reaches the specified minimum value (epsilon\_min = 0.1). This allows the model to start with a high exploration rate and hence learn quickly during the initial steps and gradually reduce exploration and increase exploitation as the model starts learning the optimal policy.

For each of the actions, the environment takes a distinct set of steps to calculate the current rewards. These Reward calculations will be described in the next section. The state returned after any action is similar and as described in the previous section ie the environment returns the current returns + historical\_n returns.

* 1. REWARDS

A **reward** is the feedback provided by the environment that measure the success or failure of an agent’s actions in a given state. For our environment, the rewards are calculated distinctly for each trading action (Buy, Hold, Sell).

The reward for each action is primarily the returns generated for each action. Further more, we have introduced the concepts of TRANSACTION COST and HOLDING COST to account for the real-life trading costs.

* Transaction cost = the cost of executing a trade and is deducted from the rewards of the Buy action. This is a % cost of the $$ stock bought at the Buy action.
* Holding Cost = Financing cost of holding a position over a Hold action step and is deducted from the reward of the Hold action. This is a % cost of the $$ position that is Held over that step.

Hence the rewards from the environment for each action are as follows:

* Reward for BUY action = Transaction cost of the stock bought
* Reward for HOLD action = Holding cost of the current position
* Reward for SELL action = Profit or Loss from selling the open position.

In order to generalize the agent behaviour for different trading situations, the environment regularizes the rewards by clipping at +/- 1. This ensures that the environment does not excessively reward/penalize an action during episodic periods of brief but excessive market volatility.

In addition, the agent (not the environment) also incorporates a discounting function (gamma = 0.97) to recognize the diminishing value of future rewards and adjust for this in the learning model.

* 1. user interface

One way for users to try out the trained model in actual trading scenarios is to use the wrapper python script (Refer to Appendix xx for user guide and demonstration). The script starts by asking user to input historical data for a selected stock. Once completed, the program will run the model and apply actions to the historical data and simulates the profits made based on the environment.

After which, user is prompted to input current date and stock information. Based on the new information, the action value will be generated for each step and the action with the highest action value is recommended to the user.

The program will also advance the model by applying the action to calculate the reward and profit of the next state.

The current state of the program only runs based on a model which only requires user input of the closing price. Once there are other models generated using other technical analysis, the program is easily scalable to include user inputs for that information.

1. model evaluation

Is having a strategy that can provide us with a profit good enough for us to say that the model has been successful? If not, what metrics and benchmarks do we have to use to compare the performance of the strategy and consequently the performance of the reinforcement learning model? This section will explore the metrics to be used and use those metrics to compare the performance of our strategy.

* 1. Metrics used for evaluation

“The riskier the investment the higher the returns” is an adage that still holds true very much today. Similarly, our metrics used to evaluate our strategy has to consider the balance between the risks and rewards. To this, we propose 2 different benchmarks that we can use to compare the returns of our strategy.

**Risk-free returns**

Investment returns other than rewarding the undertaking of additional risk, also has a function of rewarding the opportunity cost of investment. Risk-free returns in way signifies the minimum returns that strategy needs to provide to cover the opportunity cost of making that investment, notwithstanding the risk that is undertaken by the strategy. Risk-free returns are typically estimate by using the treasury bonds issued by countries with highly rated credit ratings. In our case, we will be using the US 10-year treasury bond rates.[[1]](#footnote-1)

**Expected returns**

To understand the expected returns to take on a particular amount of risk, we refer to capital asset pricing model, more commonly known as the CAP-M model.

In the CAP-M Model, other than risk free returns, expected returns also depend on market returns and beta. Market returns measures how well the market is performing while, beta measures how much additional volatility does the stock have against the market, which is also the systemic risk of the stock.

The intuition behind the CAP-M model, is that the investors are rewarded based on the additional systematic risk that taken vis-à-vis the market. This also suggests that unsystematic risk should not be rewarded as the risk can be mitigated by diversification. Our reference for the performance of the market will be the Vanguard S&P 500 Exchange Traded Fund, which mimics the S&P 500 with an accuracy of 99.7%.[[2]](#footnote-2) A 5-year monthly beta is used in this case.[[3]](#footnote-3)

**Strategy returns**

The returns from the strategy will be the pure profit or loss derived by the strategy excluding transaction and holding cost. Although, transaction and holding cost are part of the rewards of the reinforcement learning model, it is not included in the returns of the strategy. This was done to ensure that the returns of the strategy is comparable to the two benchmarks that we have proposed, as transaction and holding cost are also not factored into the benchmarks.

**Maximum drawdown**

Although, risk-free and expected returns are good benchmarks to compare the returns of the strategy, it will also be important to understand the maximum risk that is undertaken by the strategy. This is especially important as the principal in the investment in stocks are not protected. Therefore, in addition to the returns benchmark, we propose an additional metric of maximum drawdown. Maximum drawdown refers to the maximum loss (mark-to-market) that is incurred by our strategy based on the current portfolio holdings.

* 1. training results between iterative improvements

As discussed in Section 2, an initial reference code was taken from Kaggle to model the environment. However, as the code was more of a proof of concept, we found many areas that can be improved upon to make the model more realistic. This was performed iteratively to achieve the final deep Q learning model described in the Section 2.

* + 1. Improved base model

The improved base model includes the various improvements that was made to initial reference code, including the rewards functions and training logic, among others.

Initially, the rewards per time step was clipped, with any positive profits resulting in a +1 reward, while any negative profits resulting in a -1 reward. We felt that this was not indicative of the returns made by the strategy and equated the profits/loss as the rewards.

The states of the model consist of previous prices of the stock and initialised with zeros. In the initial model, the prices of the stock are be added at each time step during training, resulting in many states in the memory without previous prices, especially at the start of each episodic run. We felt that this was not indicative of the real environment, as most stock will have their prices available and can be used for prediction. Therefore, we ensure that the previous prices are fully initialised before adding the transition to the memory buffer and used for training.

In the improved base model, we only made use of the closing price of the profit/loss of the portfolio and the closing price of the past 90 days as the states for model training. Transaction cost and holding cost is set to 0.

The results of the improved base model are as shown. The cumulative profit of the strategy is almost equal to the expected stocks returns calculated from CAP-M for most of the period with some out performance in Q2 2017. However, it does outperform the risk-free interest rates. The strategy also had a maximum drawdown of USD $623 with an average portfolio size of USD $3365, signifying a 18% dip in average portfolio size in its lowest performing times.

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Graphical user interface, chart

Description automatically generated with medium confidence

* + 1. Phase 2 – Adding technical analysis

To improve on the predictivity of the reinforcement learning model, additional technical analysis variables are added states variables are added to the as state variables. It includes indicators such as trade volume, moving averages, bollinger bands and stochastic oscillator. As this are typical technical analysis variables used by traders in the financial markets, the intuition that the additional of such variables allows for the better derivation of the action values and consequently the best actions. Transaction cost and holding cost is set to 0.

The results of phase 2 with the addition of technical analysis indicators are as shown. The cumulative profit of the strategy outperformed both expected stocks return calculated from CAP-M and risk-free interest rates. The strategy also only a maximum drawdown of USD $222 with an average portfolio size of USD $1284, signifying a 17% dip in average portfolio size in its lowest performing times.

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

* + 1. phase 3 – adding transaction and holding cost

To prevent the model constantly buying and selling stocks without the consideration of transaction cost, a 1% transaction cost is added to the model. Similarly, to consider the impact of opportunity cost of holding a portfolio a daily holding cost of 0.1% is added to discourage excessive holdings and promote efficient cash management strategies.

The results of phase 2 with the addition of transaction and holding cost are as shown. The cumulative profit of the strategy outperformed both expected stocks return calculated from CAP-M and risk-free interest rates, albeit just slightly for against expected returns. The strategy had a maximum drawdown of USD $129 with an average portfolio size of USD $1120, signifying a 11% dip in average portfolio size in its lowest performing times.

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Graphical user interface, chart, scatter chart

Description automatically generated

* + 1. COMParison between different iterations

No surprise phase 2 does better than phase 1 as there are more state variables for prediction.

Phase 3 does worst then phase 2. No surprise here also as it is difficult to beat the market with

The model is not stable.

1. conclusion

**Appendix:**

1. https://www.macrotrends.net/2016/10-year-treasury-bond-rate-yield-chart [↑](#footnote-ref-1)
2. https://personal.vanguard.com/pub/Pdf/sp968.pdf?2210142649 [↑](#footnote-ref-2)
3. https://sg.finance.yahoo.com/quote/GOOG/key-statistics?p=GOOG [↑](#footnote-ref-3)