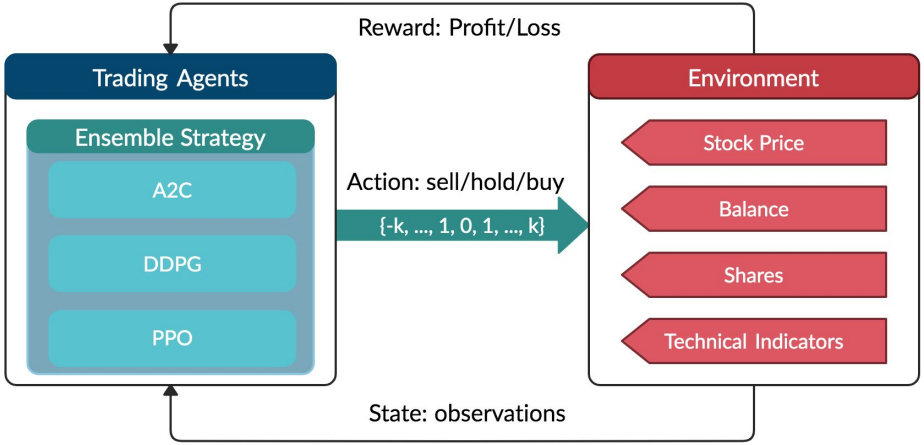
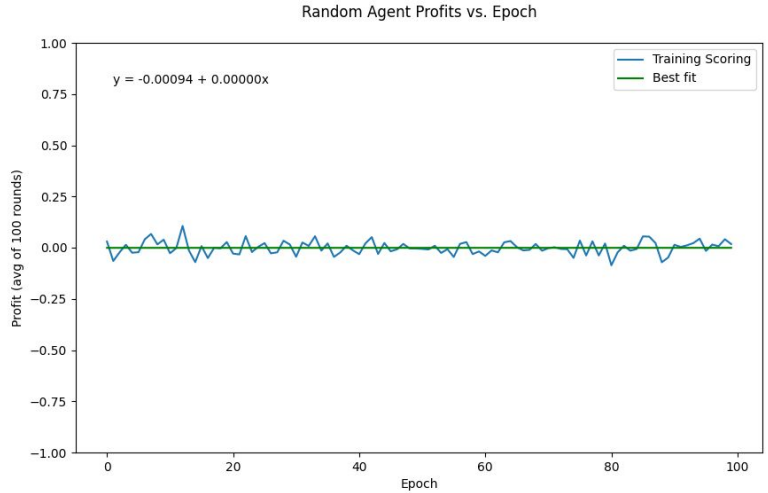
Reinforced Learning of Cryptocurrencies:
A Study in RL Techniques for Stock Market performance

By: Colton Hill

Project Summary

- **Problem:** stock trading is complicated, and crypto is worse, chaotic
- **The goal:** to document the process of creating a crypto trading RL agent
- **The plan:** Use RL to find a potential pattern in the price history *alone*





Environment

- **Master Price History**: prepossessed raw price history striped of all technical indicators and metadata.

- **State Representation**: Subsequence of the Master Price History, sized to last x steps (often 128). This subsequence is normalized and randomly scaled to help overfitting.

Agent

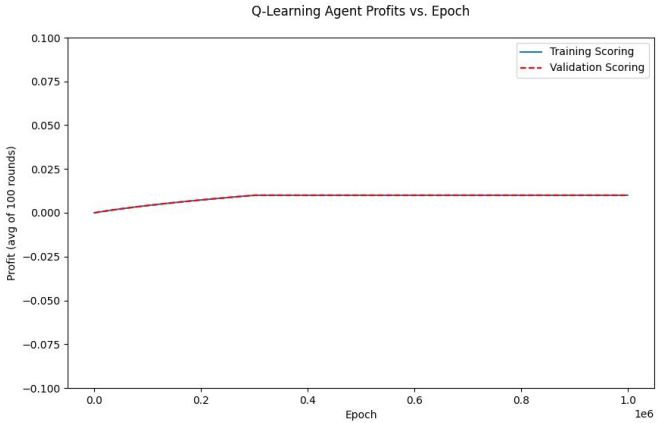
- **Action space**: sell/pass(0) or hold/buy(1) preferences. This recused the size of the state/action pairings stored by the agent.

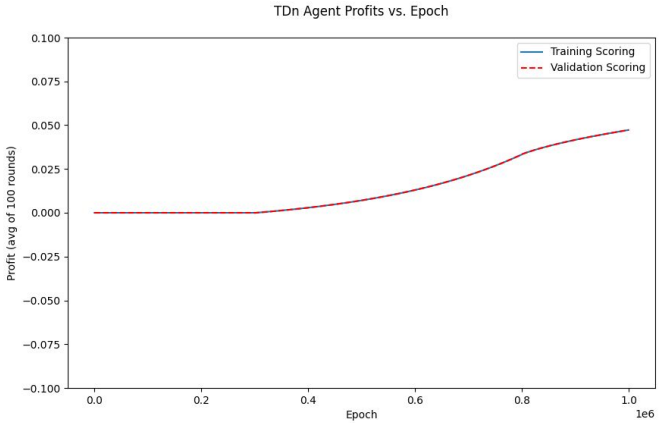
- **Reward**: The reward is simply the net profit made by the agent (since last buy/sell cycle).

- Types: Random, Q-learning, TDn, NN, Gym-wrappers(A2C,PPO)

Agents

- Q-Learning:
  - Uses bins of normalized prices (0-0.1,0.1-0.2,etc).
  - Each bin contains a favorability for each action.
  - Plateaued at 1% profit
- TD(n):
  - Uses ∆price bins (-2, 2) with 1000 bins.
  - Uses memories to retroactively distribute rewards from a sale back to an initial “buy” action.
  - Plateaued at 5% profit.

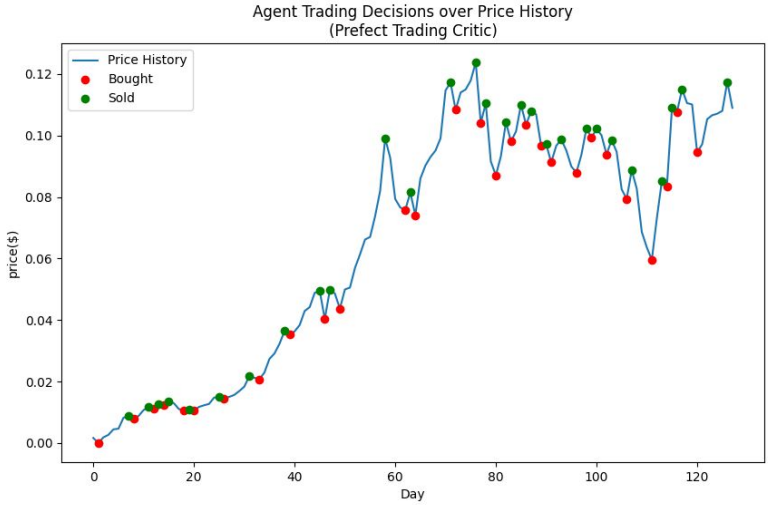


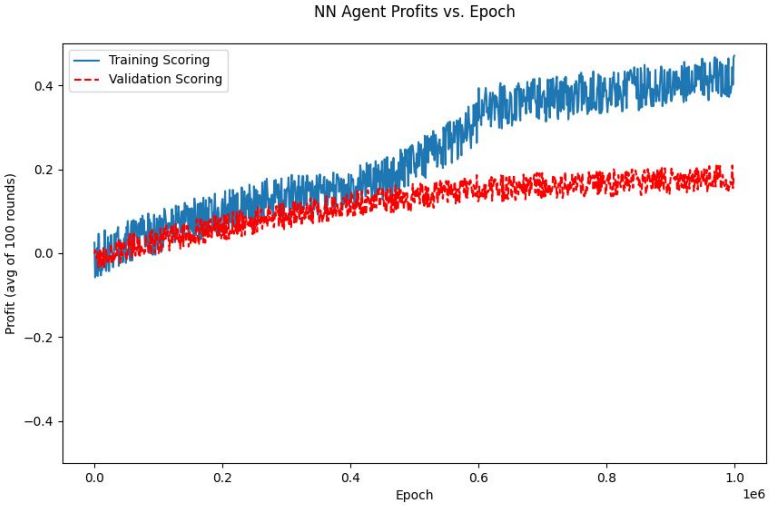


Agents

- NN-Actor/Critic:
  - Fully Connected NN using TF
  - 5 layers, 2^9→2^5, ReLu, SELU, & 0.1 Dropout on each layer.
  - Adam Optimizer, binaryCrossEntropy
  - Dynamic learning rate (~annealing)

  - Used a “perfect” (*prescient*) critic to mass generate data for batch training.
  - Still very prone to overtraining and exploiting metadata.
    - Implemented random scaling, extended data sets, dropout, all to combat this score divergence.
  - Plateaued at 15% profit





Agents (3rd-party)

- Gym-wrapper Agents
  - For my own sanity, I attempted to use 2 pre-defined agents from “stable-baselines gym-anytrading”
  - Both of each however were designed to work *WITH* the technical indicators.
  - For reference, these agents were able to on average make returns of 40% on tradisional stocks (according to the dev sight).

- A2C:
  - Ran over night (~6).
  - Used pretrained model
  - Plateaued at 9% profit.

- PPO:
  - Ran over night (~6).
  - Used pretrained model
  - Plateaued at 11% profit.

Discoveries & Results

- **Crypto does have inherent (if weak) patterns in its price history!**
  - TD(n) & NN both show positive returns.
  -
- **NN-Problems with Metadata:**
  - Metadata slipping in is a nightmare
  - Rewarding is difficult
  - Critic choice is not settled

Final Evaluation($):	
Baseline:	0%
Q-table:	1%
TD(n):	5%
NN-actor/critic:	15%
Gym-A2C:	9%
Gym-PPO:	11%

Summary/Conclusion

RL can learn to beat the market:

- Though the NN agents tend to be the best, other “simpler” methods can find success.
- Current price does correlate with depend on previous the price history.
- Not recommended, very volatile *(don’t try this at home)*