Soft-Actor Critic for Strategy Agnostic Hourly Trading

Ethan Reinhart

05-28-2025
Oregon Quant Group

Agenda

OQG
OREGON QUANT GROUP

01 Soft-Actor Critic Architecture

02 Data Aggregation

03 Feature Engineering

04 Autocorrelation

05 Reward Shaping

06 Results

Thesis

OQG
OREGON QUANT GROUP

Reinforcement Learning provides the capability of an agent to execute actions based on observed sequences to maximize its reward.

Training an agent on hourly stocks prices of curated tickers combines predictive ability of **Deep Learning** with entropy and dynamic, policy-ambiguous decision making.

A traditional alpha reward allows for **ROI maximization** with **dynamic regime** switching based on learned experiences.

Soft-Actor Critic

OREGON QUANT GROUP

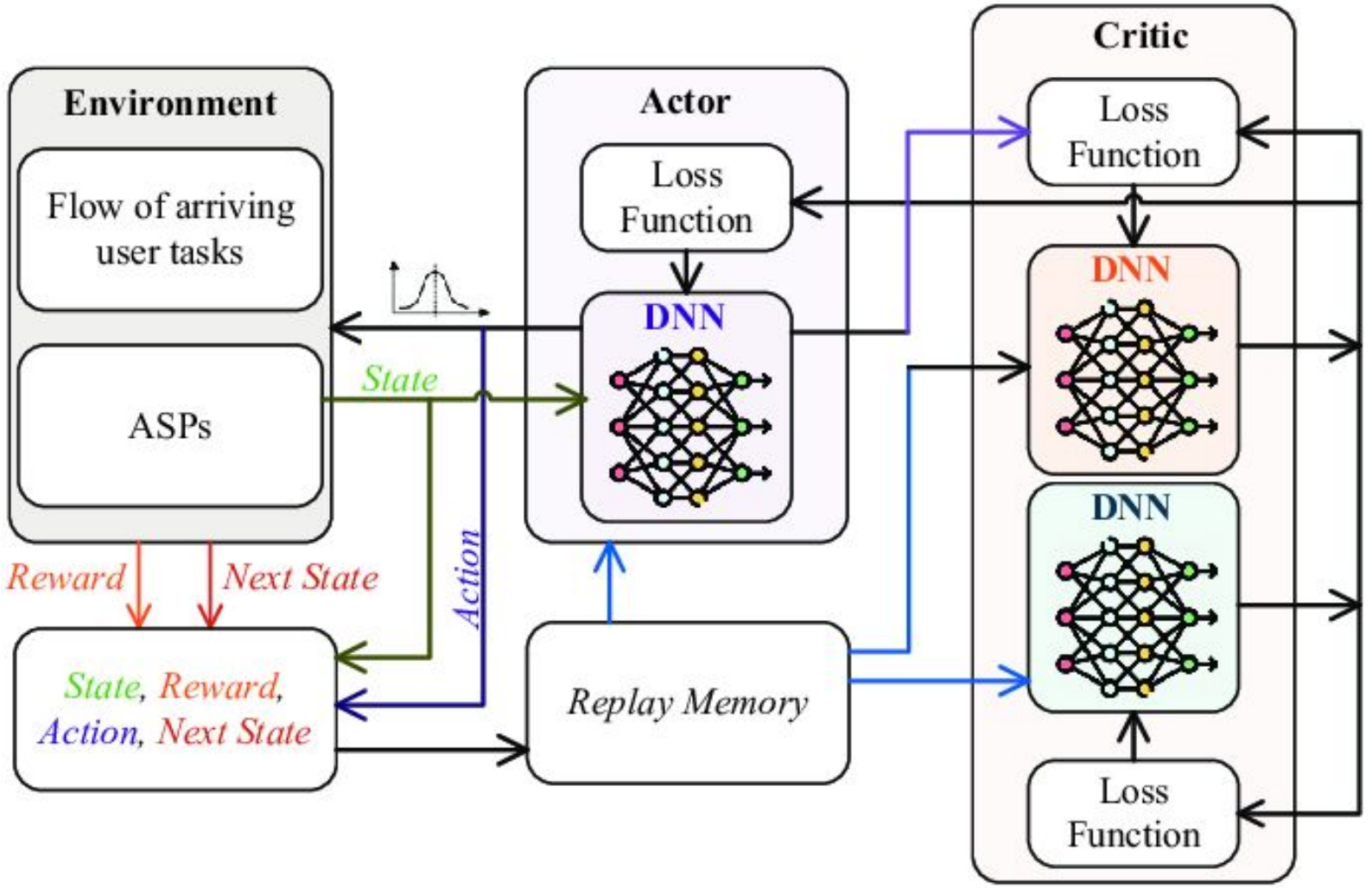
1. Reinforcement Learning
2. Different from Q-Learning (single action vs discrete action space vector)
3. Introduces entropy balance to force exploration
    a. Entropy: state disorder
4. Off-Policy actor critic - updates from replay buffer after each episode
5. Basic premise:
    a. Critic predicts expected value of taking action in current state
    b. Actor maximizes expected value + entropy
        i. Max reward with max exploration (great for avoiding local minima)

Soft-Actor Critic

OREGON QUANT GROUP

1. **Replay Buffer**
   a. Stores reward, action, and state transition
   b. Updates actor and critic with batch sampled updates
2. **Twin Q-networks (Critics)**
   a. Jointly trained Q-networks, prevents bias by taking minimum of both
3. **Target Q-networks**
   a. Slowly updated copies of the two critic networks that are updated with Polyak averaging
   b. Tau controls update: $\overline{\theta_i} \leftarrow \tau\theta_i + (1-\tau)\overline{\theta_i}$
4. **Actor (Policy)**
   a. Alpha = entropy constant, pi(a|s) = randomness of this action given Gaussian policy distribution $J_{\text{actor}}(\phi) = \mathbb{E}_{s_t \sim D, a_t \sim \pi_\phi}[\alpha \log \pi_\phi(a_t|s_t) - Q_{\min}(s_t, a_t)]$
   b. Q(s,a) = expected value of action in current state

Soft-Actor Critic



Data

1. Alpaca API (Best source I could find)
   a. Open, Close, High, Low, Volume, Volume Weighted
2. Data: 9 years; 7 hours per day
   a. 9:30, 10:30, 11:30, 12:30, 1:30, 2:30, 3:30 (ET)
   b. 251 * 9 * 7 = ~15500 data points per ticker
3. Tickers: U(NASDAQ + S&P)
   a. *Project not finished so only used: ["AAPL", "MSFT", "A", "AMD", "JPM"] for testing
4. Grows proportional to tickers → tickers observed by each agent must be controlled
5. Stored in local Postgre DB for fast read

Features

OQG
OREGON QUANT GROUP

1. Started with prices and returns for each stock
   a. Sub-par performance, would stay about even or slightly negative, often converged to not trading as best action
   b. Why might this be?
4. **Efficient Market Hypothesis**: asset prices reflect all publicly available information
   a. Consequently, hourly/daily changes in stock prices should behave like white noise
2. Proof? Autocorrelation →AR(1) < 0
3. More meaningful information can be derived using stats

Autocorrelation

OQG
OREGON QUANT GROUP

1. **Autocorrelation** is mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals.

$$\rho = \frac{\sum_{t=k+1}^{T}(r_t - \overline{r})(r_{t-k} - \overline{r})}{\sum_{t=1}^{T}(r_t - \overline{r})^2}$$

2. AR(n) process autoregressive process is one in which the current value is based on the immediately preceding value n values (AR(1) based on previous)
3. **Formulas:**
   a. Window Size = take first AR(n) such that AR(n) < epsilon (I used 0.1)
   b. Essentially most recent timestep where the window reveals 90% relevant info
   c. Gamma (future look-ahead for agent reward) = $0.5^{\frac{1}{h}}$

$$h = \frac{\log 0.5}{\log AR(1)}$$

Feature Engineering

OREGON QUANT GROUP

1. 5-Hour Momentum: $\frac{p_t - p_{t-5}}{p_{t-5}}$
2. Moving Average Crossover: 5hr SMA - 20hr SMA
3. Volatility: Rolling Std of Returns
4. Relative Strength Index: $100 - \frac{100}{1 + \frac{14hr(+)\Delta SMA}{14hr(-)\Delta SMA}}$
5. Normalized Volume: $\frac{V_t - \overline{V}}{\sigma}$
6. Returns: raw %

Hyperparameters

OQG
OREGON QUANT GROUP

1. Slice length, batch size, network size, buffer size, eval frequency, gradient steps, tau, gamma, learning rate, clip range, alpha, lambda, number of tickers observed
   a. Picking just 2 for each: 8192 trials required
   b. Training per trial takes around 15 min on my 4090
   c. About 2048 hours conservatively
   d. About 85 full days
   e. Impossible in a term→I had to guess and make assumptions so mine are by no means optimal
2. Custom Walk-Forward hyperparameter tuning implementation
   a. Define train, eval, and test size; eval was my step size
   b. Train on [0, train], stop on [train, train + eval], test on [train + eval, train + eval + test]
   c. Generically: [step * i, train + step * i]; [train + step * i, train + eval + step * i]; [train + eval + step * i, train + eval + test + step * i]
   d. Forces hyperparameters to be generalizable

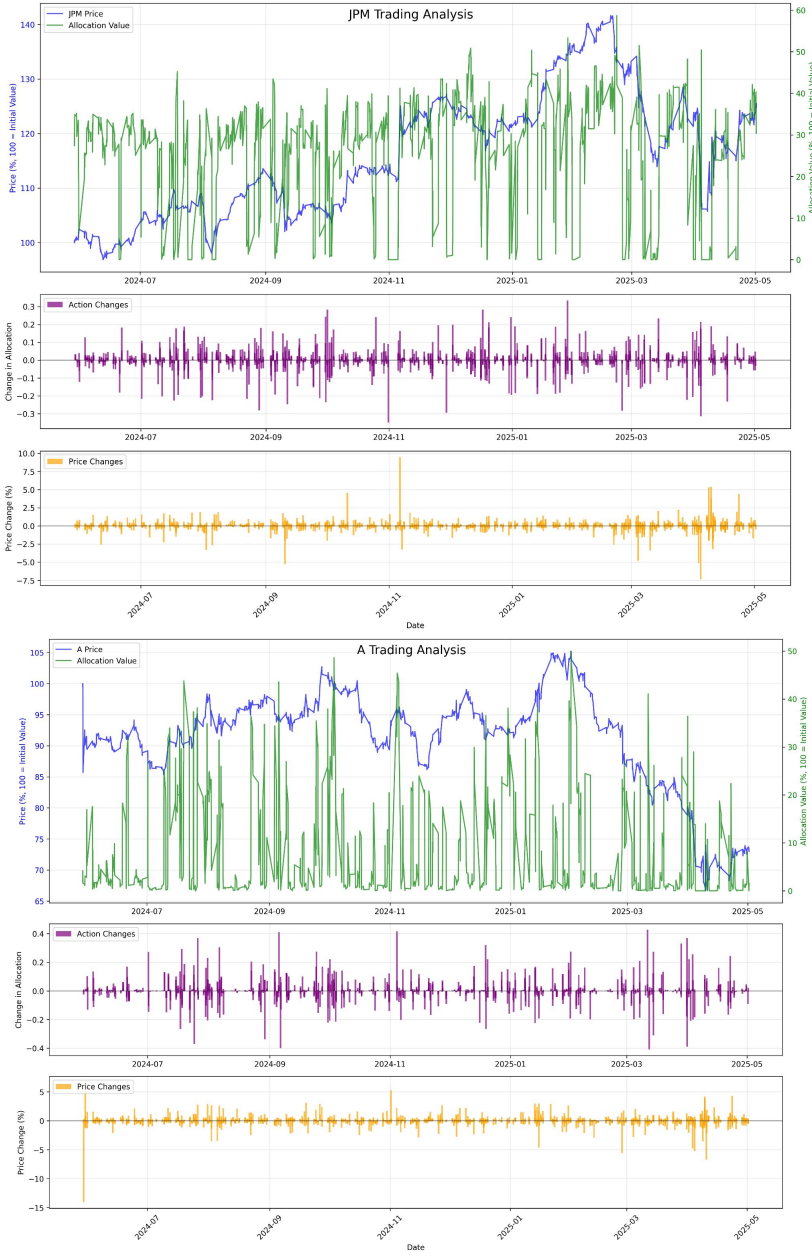
Reward Shaping

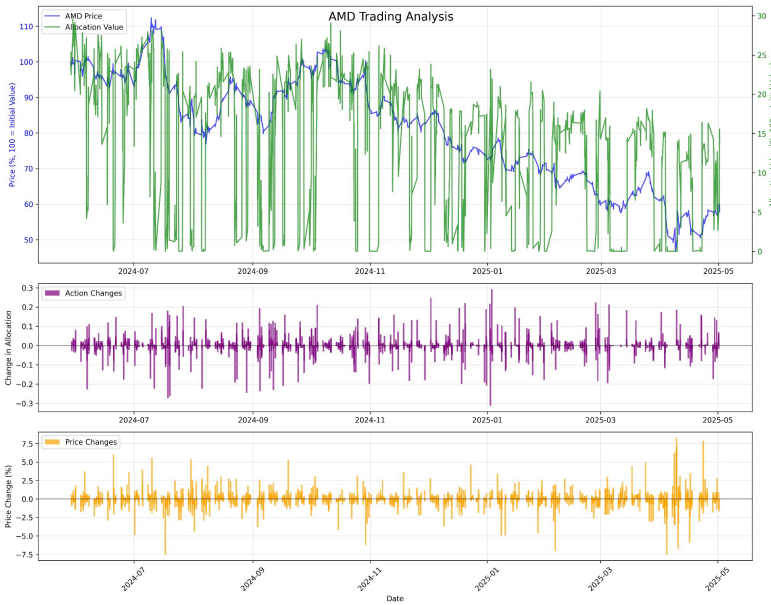
OQG
OREGON QUANT GROUP

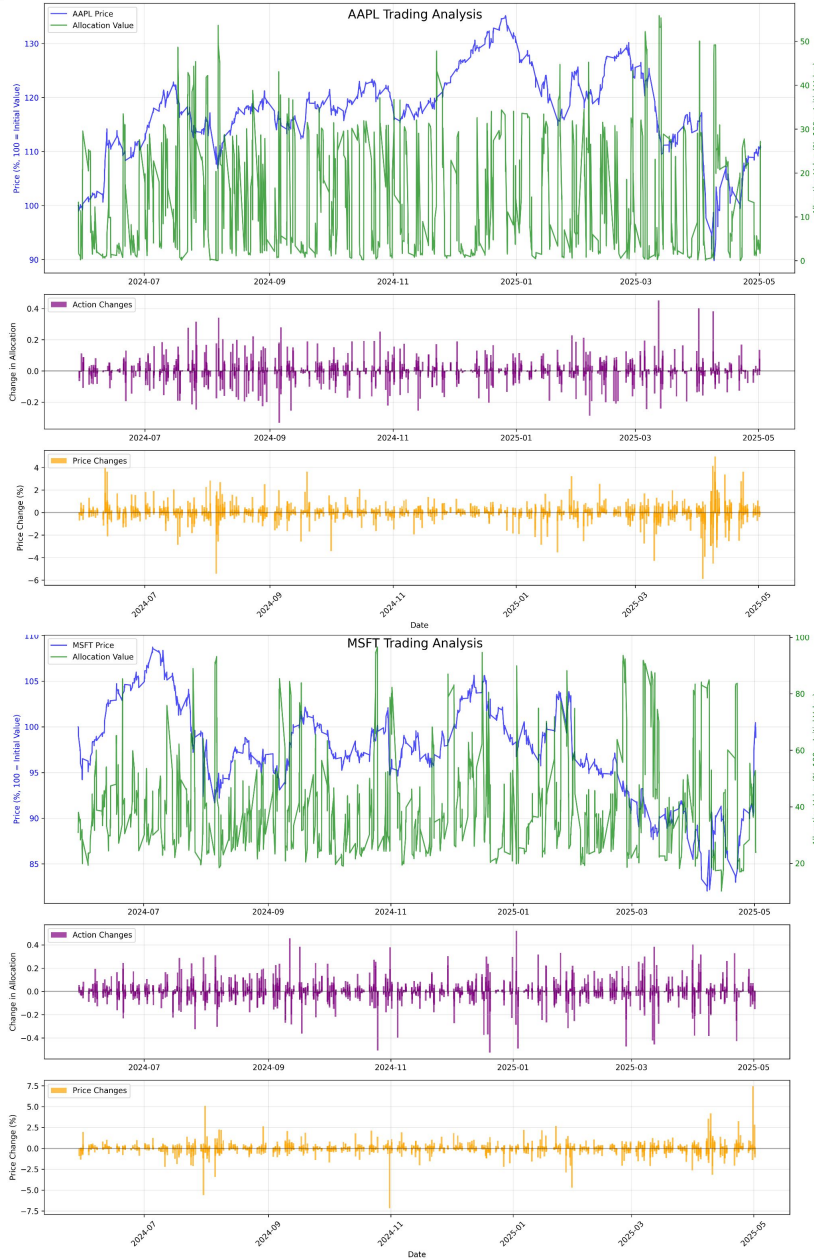
1. How to reward?
    a. Portfolio Return (including trading fees)
    b. Return Compared to Benchmark
    c. Risk Penalty for high volatility
2. Reward clipping for stable gradients
    a. **#1 Consideration**: Summed clipped gradients must be positive (want smallish)
        i. Negative summed gradient clip means you are ignoring more large losses
        ii. Positive summed gradient clips means you are ignoring more large wins
    b. Better to miss profit than have large drawdown

Key Finding

OQG
OREGON QUANT GROUP



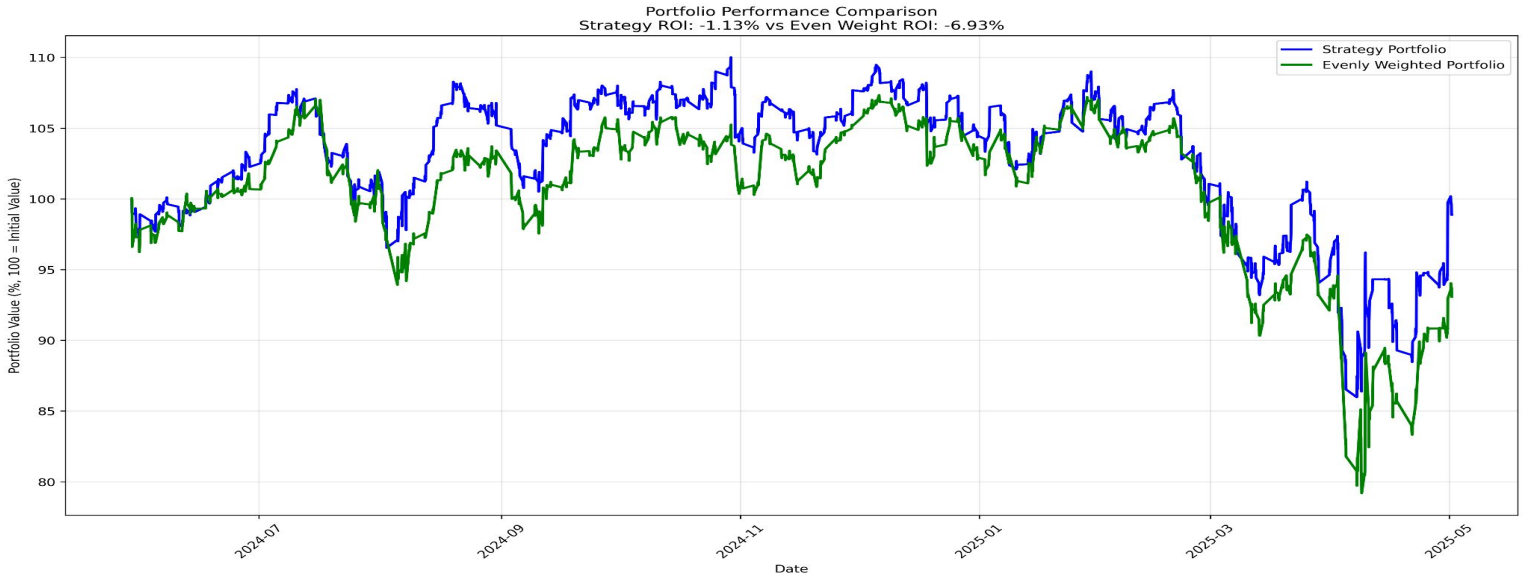




Key Finding

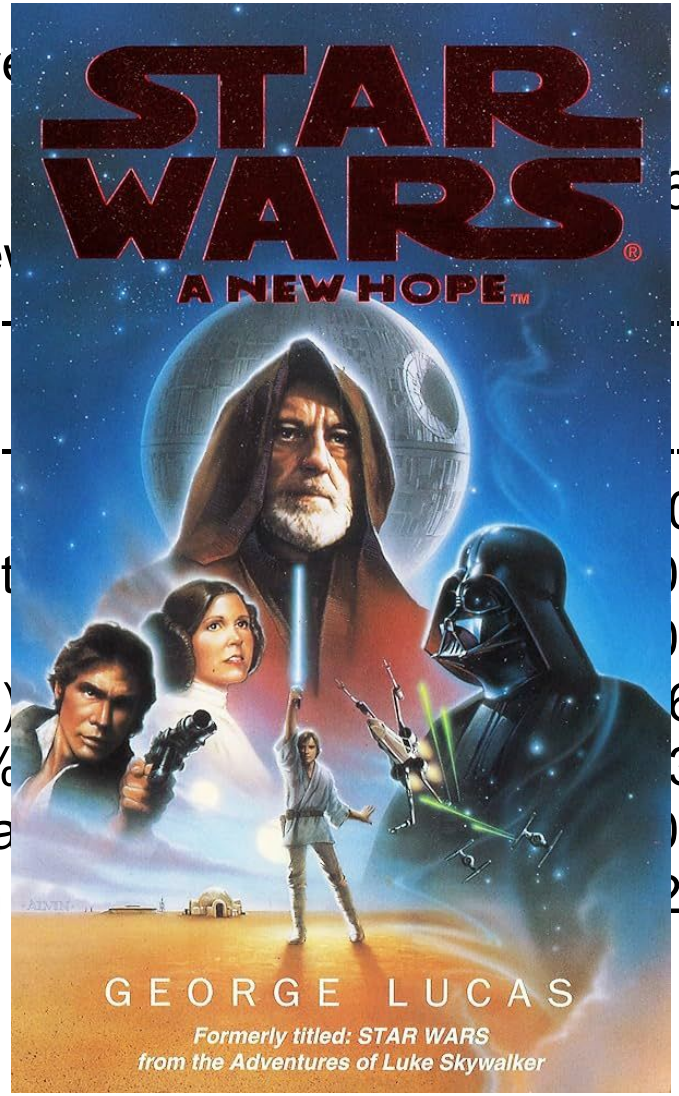
OQG
OREGON QUANT GROUP

	My Agent	Even Diversification
Net ROI %	-1.13	-6.93
Avg Hourly Return %	<0.001 & >0	>-0.001 & <0
Daily ROI %	0.01	-0.02
Annual ROI %	-1.23	-7.52
Monthly ROI %	-0.10	-0.65
Annualized Sharpe	0.05	-0.21
Max DD %	-21.85	-26.23



Key Finding

Tickers Observe	'AMD', 'JPM']
Slice length	
Interval	6
Total clipped rev	
	Portfolio
Net ROI (%)	04
Mean hourly ret	00
Daily ROI (%)	00
Annual ROI (%)	60
Monthly ROI (%	30
Annualised Sha	05
Max drawdown	2.59



Conclusion

OQG
OREGON QUANT GROUP

1. Efficient Market Hypothesis in Action
2. Almost impossible to beat the market with only market data
3. Planning on making more advancements in the future
    a. Transfer learning on daily data so weights are seeded with regime detection
4. Integrate with Alpaca if I can get maintainable positive over benchmark

Me developing my 5th algorithmic trading strategy after the first 4 lost me 60% of my initial investment

Thank You