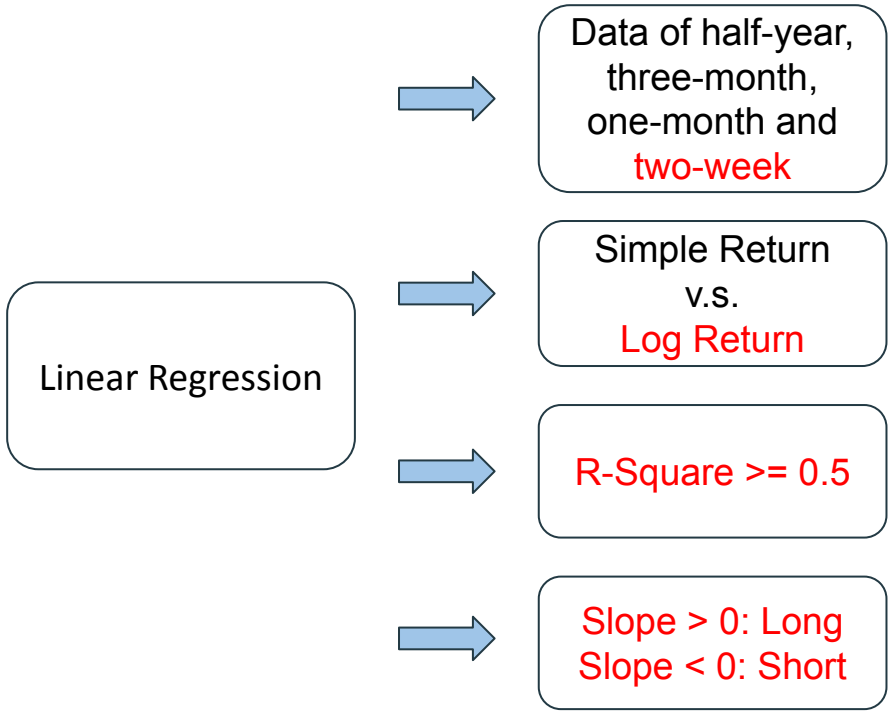
Alpha Group Final Presentation

Group Members: Yichong Huang, Zhongtian Chen, Zejia Qian, Xiance Zhang

Project Timeline

- Week 1-2: Exploratory Phase
  - Rank 0.2 + Linear Regression + Fundamental Analysis
- Week 3-4: Quantimental
  - ML Model + Fundamental Analysis
- Week 5-10: Model Finalized + Improving
  - Deep Learning Return Prediction + Optimization

Week 1-2 : **Linear Regression** and Fundamental Analysis



	slope	intercept	R_squared
EWC	0.003600	-0.011496	0.779571
JPM	0.003464	-0.010072	0.697887
OGN	0.010091	-0.049344	0.686012
XLF	0.002007	-0.005878	0.666488
AVY	0.003067	-0.014305	0.644398
PRU	0.004118	-0.010506	0.628934
ICLN	0.004352	-0.016797	0.607398
CTAS	0.004246	-0.010458	0.59429
EWU	0.002630	-0.007233	0.59344
IWM	0.003139	-0.014507	0.579887
AXP	0.002571	-0.007317	0.523834
EWZ	0.004695	-0.012672	0.518117
IGF	0.002579	-0.008286	0.510282
XLI	0.002142	-0.011215	0.500168

Week 1-2 : Linear Regression and **Fundamental Analysis**

Developed Market vs. Emerging Market

Analyst Reports News

Valuation (Undervalue vs Overvalue)

Event Driven

- **Macro:** GDP growth, unemployment rate, CPI growth, dollar index, yield spread, exchange rate, etc.
- **Equity:** ROE, EPS Growth rate, Revenue Growth Rate, PE Ratio, PB Ratio, and etc.
- **Commodity:** Oil Price, Gold Price, and etc.
- **Sentiment:** NAAIM Exposure Index, Put-Call Ratio, AAII's Investor Sentiment Survey
- **Technical:** Percentage stocks that are above 200-days moving average, RSI, and etc.

Week 3-4 : Autoregression and Machine Learning Model

- **Input Rank as Vector Variable:**

$$R_t = [r_{1,t}, r_{2,t}, r_{3,t}, r_{4,t}, r_{5,t}] \quad s.t. \sum_{i=1}^{5} r_{i,t} = 1$$

r_i can be interpreted as probability for each rank

- **Prediction Logic:**

1. Model training: input factors $X_{t-8}, X_{t-7}, \dots\dots, X_{t-1}$ and rank $R_{t-8}, R_{t-7}, \dots\dots, R_{t-1}$ to fit different models
2. Model Prediction: use $X_t, \dots, X_{t-i}$ and $R_t, \dots, R_{t-i}$ to predict R_{t+1}
3. Independent modelling for single asset

- **Back-testing Logic:** Rolling Prediction with Historical 100 Weeks' Data

Week 3-4 : Autoregression and Machine Learning Model

- **Vector Autoregressive (VAR) Model**

✓ A VAR(1) in two variables can be written in matrix form (more compact notation) as

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \end{bmatrix}$$

✓ When applying data to fit VAR, it yields an optimized lag order 'p'.

✓ Excluding momentum and macro factors from this model serves two purposes:

- Testing the time series property of relative ranks.

- Distinct momentum and macro factors can exhibit diverse lagging effects, which renders a singular 'p' less applicable across all.

- **Random Forest Regressor & XGBoost Regressor**

✓ Random Forest:

- It belongs to the ensemble learning methods and operates by constructing multiple decision trees during training and outputting the average prediction (for regression) or the mode prediction (for classification) of the individual trees.

✓ XGBoost:

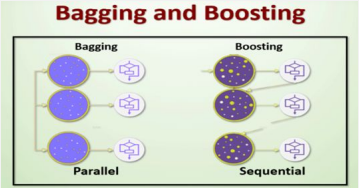
- It is an optimized gradient boosting algorithm that builds multiple decision trees sequentially, where each subsequent tree corrects the errors made by the previous one. XGBoost focuses on reducing the residuals of the previous models to gradually improve predictions.

✓ Parameter Tuning:

- Implement Grid Search CV to choose best parameter set which results in the smallest MSE of test set

✓ Avoid Overfitting:

- Early stopping being used, if the model does not see obvious improvement in MSE for 10 iteration, it would directly stop fitting

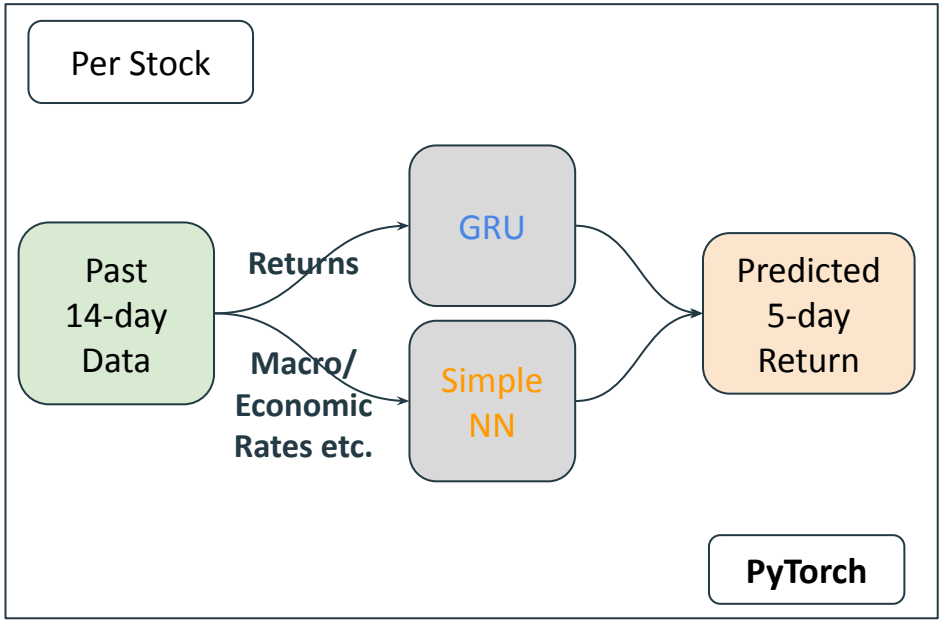


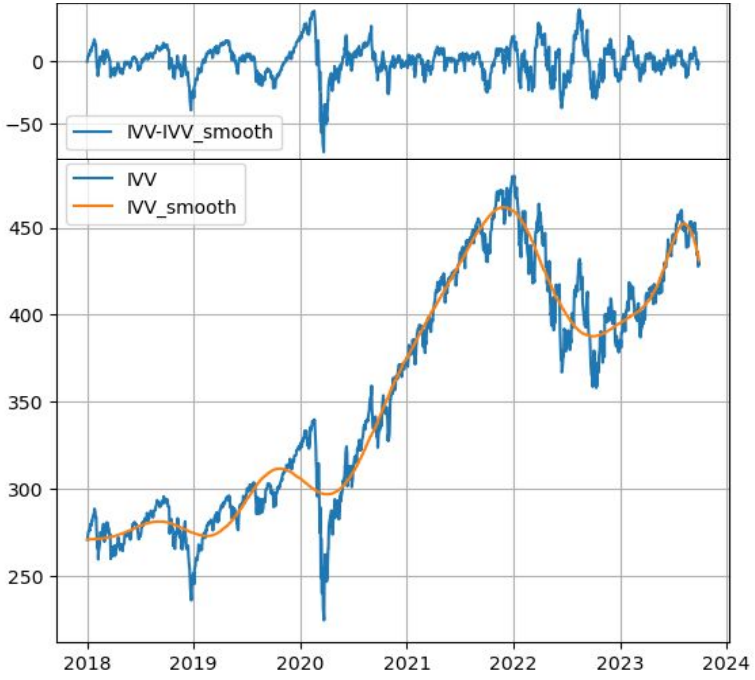
Week 3-4 : Autoregression and Machine Learning Model

- **Criterion:**
  - **FP – average predicted probability for the real rank**
  - **FSP – using softmax transformation to replace raw prediction for FP**
  - **MSE1 – mean square error for real rank and predicted rank**
  - **MSE2 – using softmax transformation to replace raw prediction for MSE1**

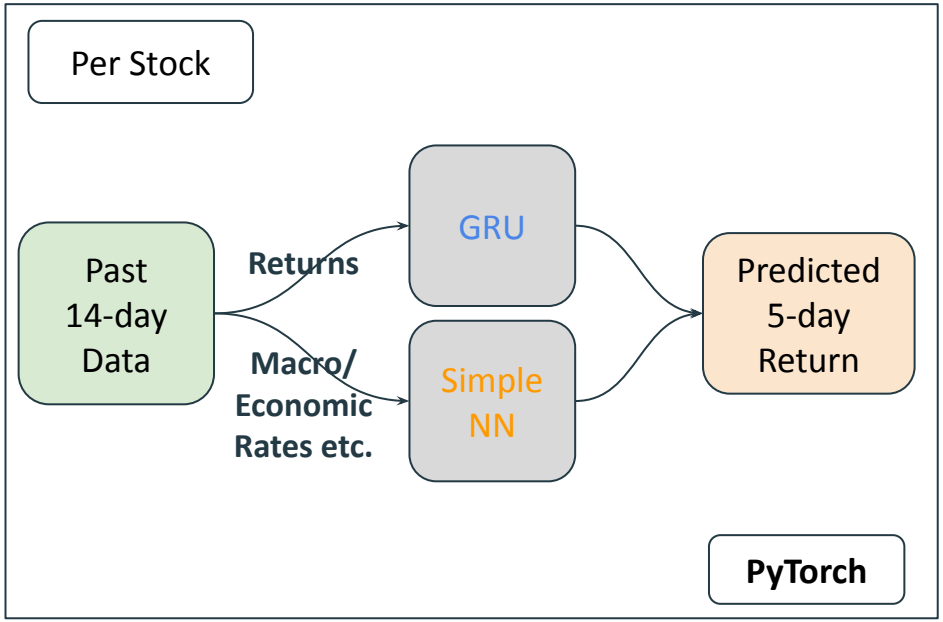
		FP	FSP	MSE1	MSE2
Benchmark - equal weight		0.20	0.20	0.16	0.16
Random Forest	median	0.27	0.21	0.17	0.17
	75 quantile	0.29	0.22	0.18	0.17
	25 quantile	0.25	0.21	0.17	0.16
XGBoost	median	0.26	0.22	0.25	0.25
	75 quantile	0.30	0.22	0.26	0.26
	25 quantile	0.24	0.21	0.23	0.23
VAR	median	0.23	0.21	0.24	0.24
	75 quantile	0.26	0.21	0.25	0.25
	25 quantile	0.20	0.20	0.23	0.23

Week 5-10: Return Prediction Model

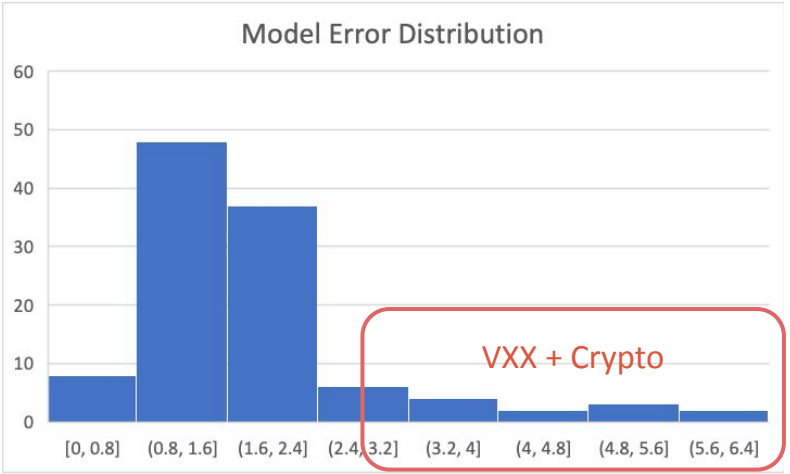




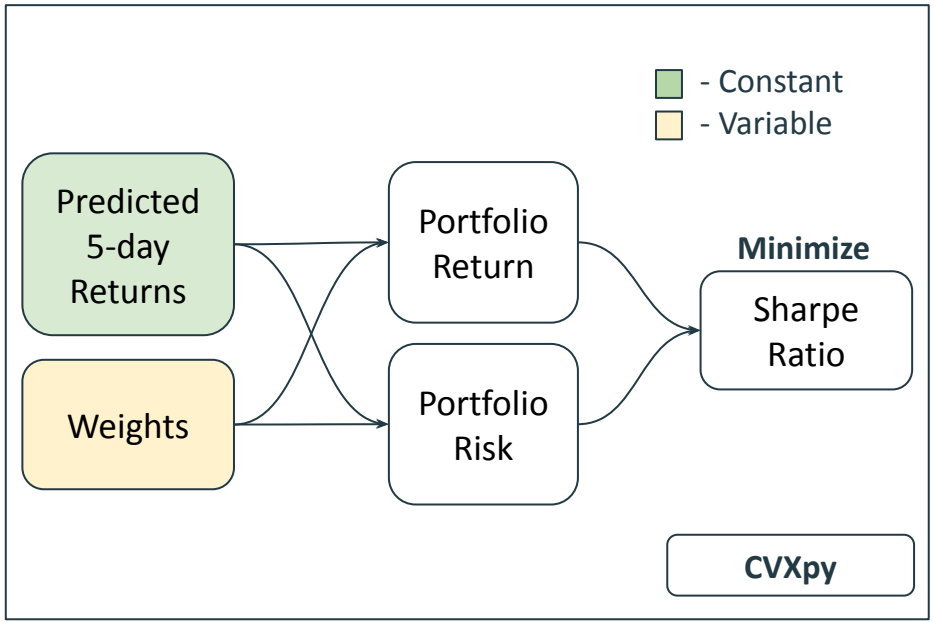
Week 5-10: Return Prediction Model



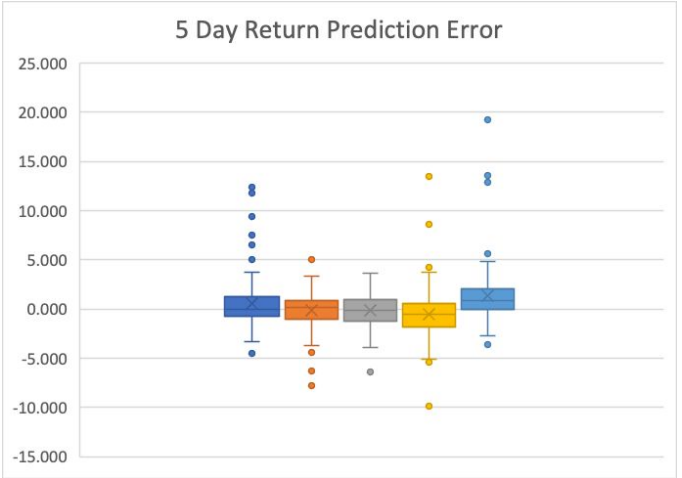
- Loss = MSE (Predicted 5-day Return, Actual 5-day Return)



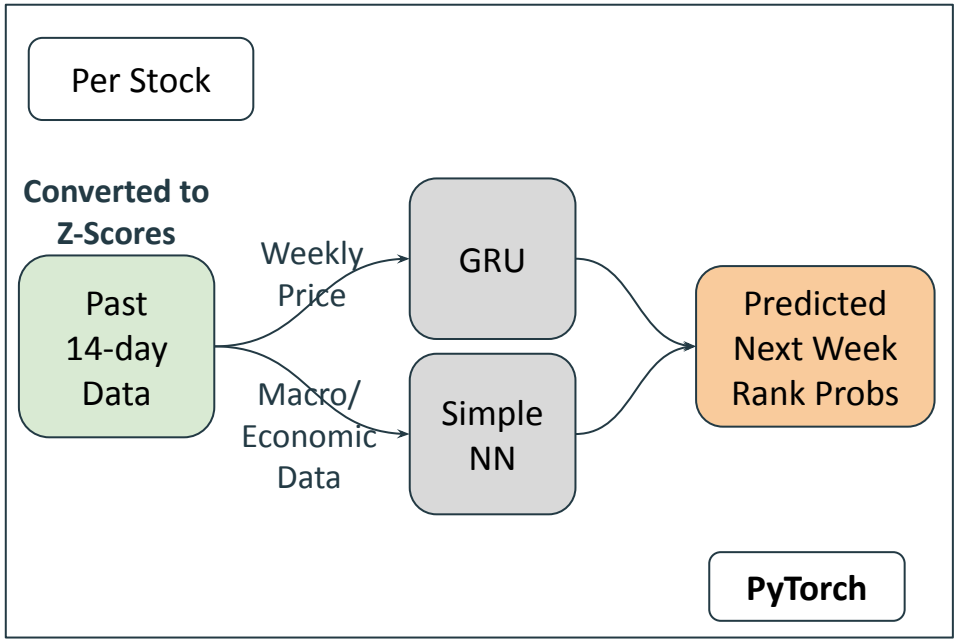
Week 5-10: Sharpe Ratio Optimization



- Very dependant on accuracy of return prediction



Also tried: DL Rank Prediction Model



- Loss = MSE (Predicted Rank Probabilities, Actual Rank Probabilities) + L2 Regularization
- Average MSE 0.11
  - *Abandoned* 👎

Results

Date	team_rank	forecast_performance	decision_performance	forecast_rank	decisions_rank	overall_rank
1-Oct	12	0.161333	-10.209063	7	15	11
8-Oct	10	0.19912	1.608041	11	11	11
15-Oct	5	0.092601	-6.600217	1	13	7
22-Oct	1	0.147793	12.42323	4	1	2.5
29-Oct	9	0.268632	22.459095	14	5	9.5
19-Nov	5	0.16	3.153181	5	11	8
26-Nov	13	0.25	4.14266	15	9	12

Thanks!