Project Presentation

Group: hello_world

Jiaheng Zhou

Yilan Guo

Tianyang Xu

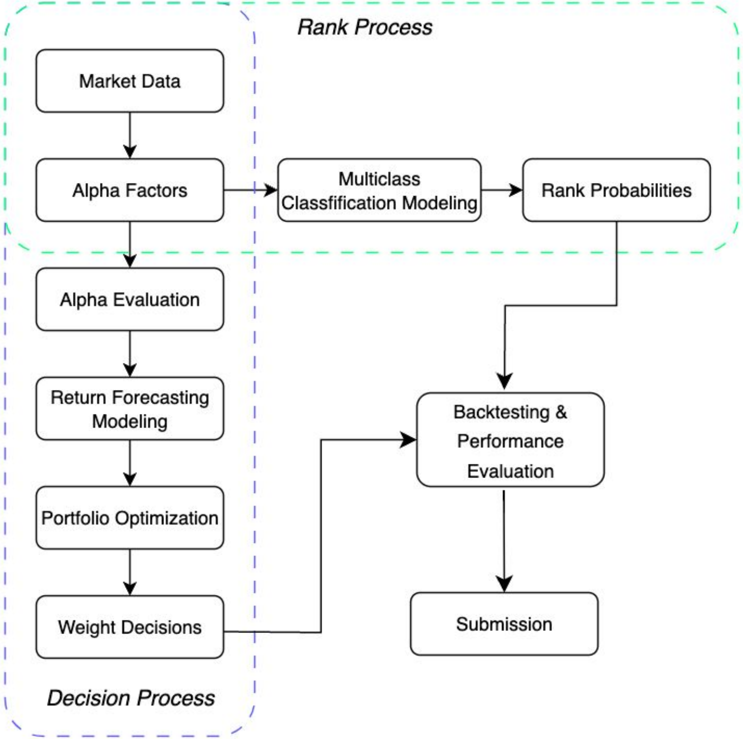
Zhengyi Zhu

Yue Fei

Dec 4th, 2023

Project Workflow & Journeys

Objective: Rank and weight decision forecasts for 110 stocks on a weekly basis.



Week 0 - 1: Random rank and decisions

Week 2 - 4: Random rank, ARIMA model to forecast returns and MVO based on historical and forecast returns for weight decisions

Week 5 - 6: Created features from market data to train on ML models and NNs for multiclass classification, weight decisions remained the same

Week 7 - 9: Performed alpha evaluations, further feature engineering, and trained ML models to forecast returns used for weight decisions by MVO

Week 10+: Focused on mining good quality alpha factors, fine-tuning the ML models, trying different optimization methods

Data Collection and Processing

Objective: collect the data, QC and resolve issues to ensure the data is of good shape on an on-going basis.

From 2020-01-01 to the latest daily data

- Market OHLC data
- Asset Class (Stock, ETF, and Crypto)
- GICS industry & sector

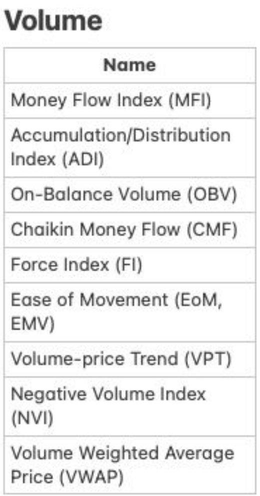
Data issues and solutions:

- Missing observations: refilling with mean/median/previous values
- Outliers: trimming or winsorizations w/ and w/o Z-score methods

Alpha Generation

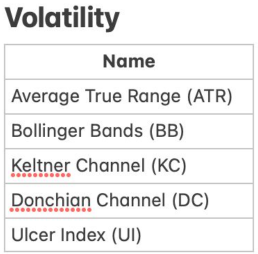
Objective: feature engineer the clean and processed data to find qualified predictive features to train models

- WorldQuant – 101 Alphas (https://arxiv.org/pdf/1601.00991.pdf)
- Technical indicators
  - Volume
  - Volatility
  - Trend
  - Momentum



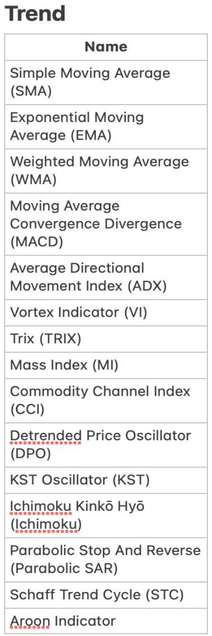
Volume

Name
Money Flow Index (MFI)
Accumulation/Distribution Index (ADI)
On-Balance Volume (OBV)
Chaikin Money Flow (CMF)
Force Index (FI)
Ease of Movement (EoM, EMV)
Volume-price Trend (VPT)
Negative Volume Index (NVI)
Volume Weighted Average Price (VWAP)



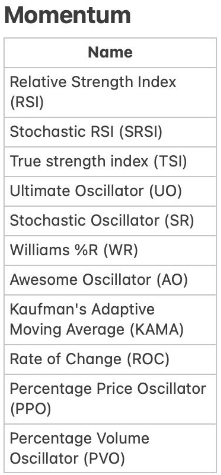
Volatility

Name
Average True Range (ATR)
Bollinger Bands (BB)
Keltner Channel (KC)
Donchian Channel (DC)
Ulcer Index (UI)



Trend

Name
Simple Moving Average (SMA)
Exponential Moving Average (EMA)
Weighted Moving Average (WMA)
Moving Average Convergence Divergence (MACD)
Average Directional Movement Index (ADX)
Vortex Indicator (VI)
Trix (TRIX)
Mass Index (MI)
Commodity Channel Index (CCI)
Detrended Price Oscillator (DPO)
KST Oscillator (KST)
Ichimoku Kinkō Hyō (Ichimoku)
Parabolic Stop And Reverse (Parabolic SAR)
Schaff Trend Cycle (STC)
Aroon Indicator



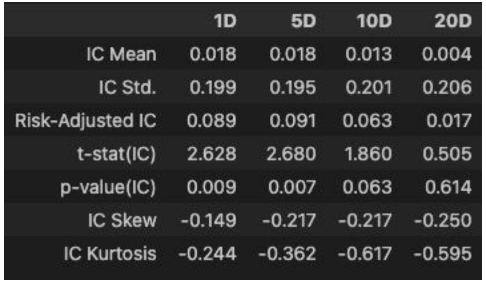
Momentum

Name
Relative Strength Index (RSI)
Stochastic RSI (SRSI)
True strength index (TSI)
Ultimate Oscillator (UO)
Stochastic Oscillator (SR)
Williams %R (WR)
Awesome Oscillator (AO)
Kaufman's Adaptive Moving Average (KAMA)
Rate of Change (ROC)
Percentage Price Oscillator (PPO)
Percentage Volume Oscillator (PVO)

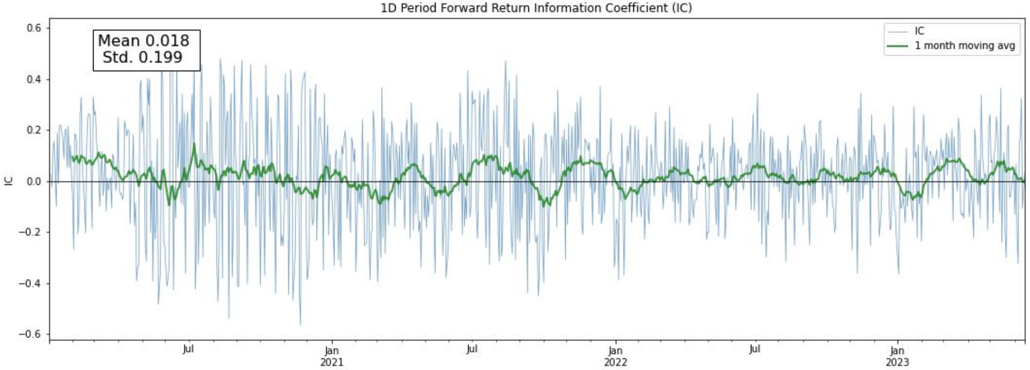
Alpha Evaluation

Objective: evaluate alpha factors and cherry-pick for predictive modeling

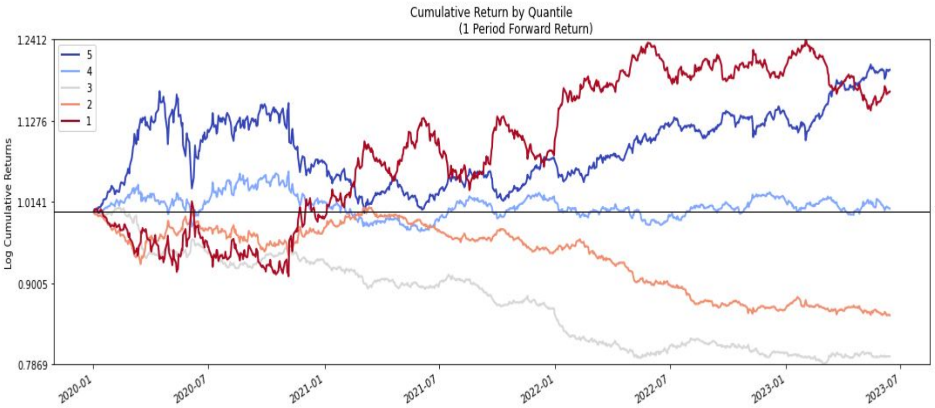
- ICIR
- Zero-investment portfolio
- Cross-sectional regression

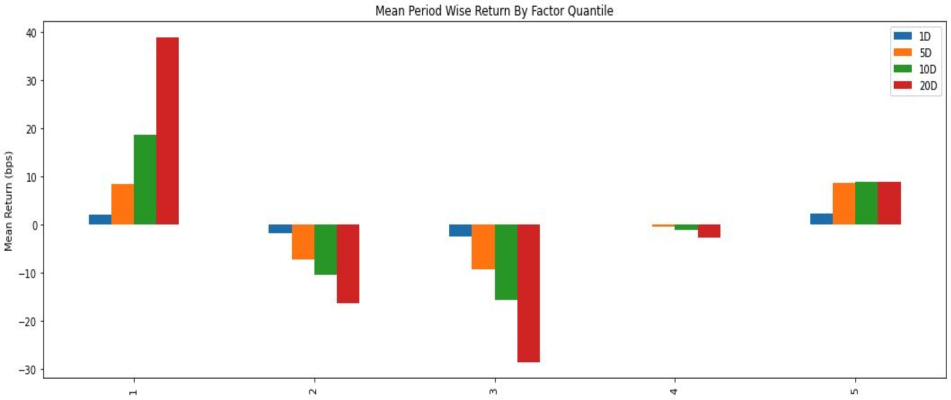


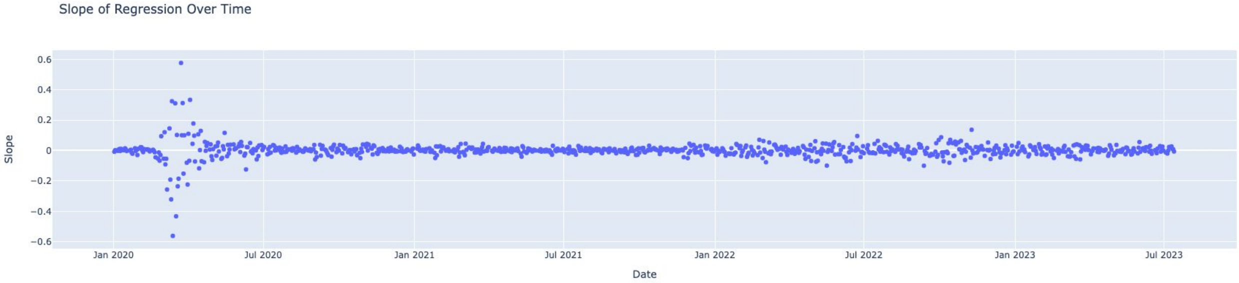
	1D	5D	10D	20D
IC Mean	0.018	0.018	0.013	0.004
IC Std.	0.199	0.195	0.201	0.206
Risk-Adjusted IC	0.089	0.091	0.063	0.017
t-stat(IC)	2.628	2.680	1.860	0.505
p-value(IC)	0.009	0.007	0.063	0.614
IC Skew	-0.149	-0.217	-0.217	-0.250
IC Kurtosis	-0.244	-0.362	-0.617	-0.595

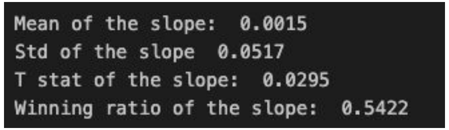


Alpha Evaluation









```
Mean of the slope:  0.0015
Std of the slope  0.0517
T stat of the slope:  0.0295
Winning ratio of the slope:  0.5422
```

Predictive Modeling

Objective: use selected alpha factors to train ML models on a multiclass classification problem for rank forecasts, and a regression problem for return forecasts, hyper-params fine-tuned with time series CV method.

Multiclass Classification

Models experimented: GLM, Decision Tree, Random Forests, Boosting Trees (LightGBM, XGBoost), Neural Networks, LSTM.

Regression

Models experimented: ARIMA, GLM, Decision Tree, Random Forests, Boosting Trees (LightGBM, XGBoost), Neural Networks, LSTM.

Finally, we chose LightGBM for both problems given considerations on the efficiency and performance.

Portfolio Optimization

Objective: construct the optimal portfolio that can produce good long-term risk-adjusted portfolio returns.

Methods

- Equally Weighted
- Cap Weighted
- Mean-Variance Optimization
- Maximize Sharpe Ratio Optimization (Last chosen)

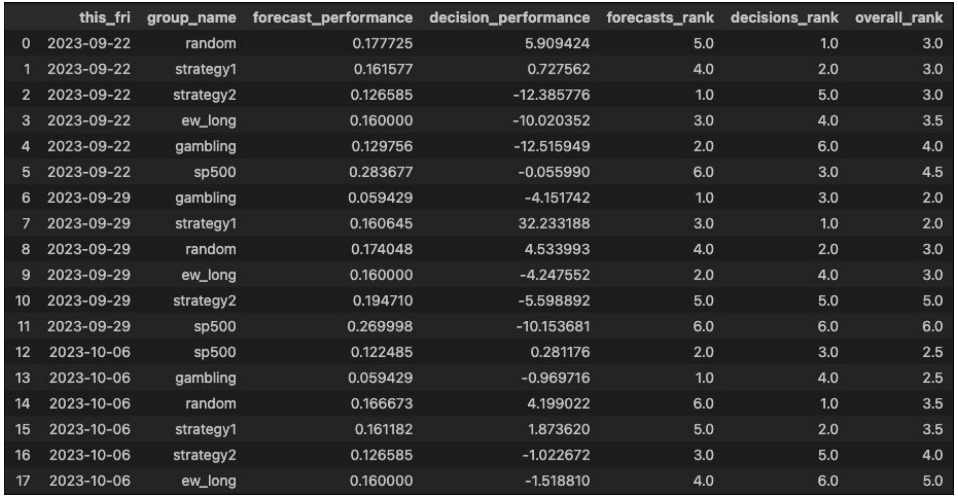
Constraints

- Long-only
- Long-short (Last chosen)

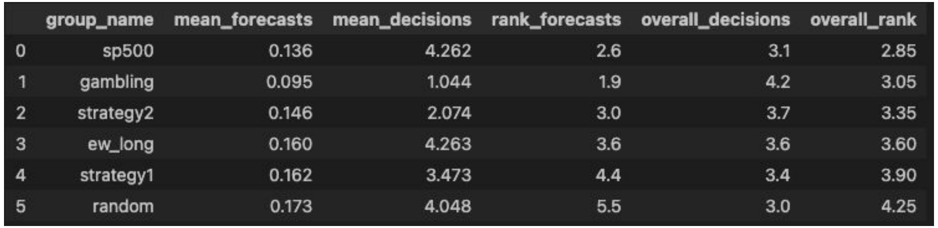
Backtesting

Objective: verify if the ultimate historical performance would beat against 4 provided benchmarks and among other solutions both in the long-term and short-term.

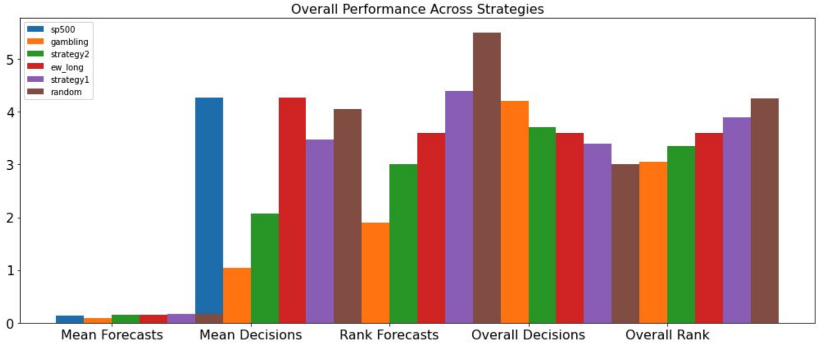
We developed a backtesting pipeline to check the historical performance per proposed strategies, for example, below are the two strategies' performance evalution, WoW and the overall.



	this_fri	group_name	forecast_performance	decision_performance	forecasts_rank	decisions_rank	overall_rank
0	2023-09-22	random	0.177725	5.909424	5.0	1.0	3.0
1	2023-09-22	strategy1	0.161577	0.727562	4.0	2.0	3.0
2	2023-09-22	strategy2	0.126585	-12.385776	1.0	5.0	3.0
3	2023-09-22	ew_long	0.160000	-10.020352	3.0	4.0	3.5
4	2023-09-22	gambling	0.129756	-12.515949	2.0	6.0	4.0
5	2023-09-22	sp500	0.283677	-0.055990	6.0	3.0	4.5
6	2023-09-29	gambling	0.059429	-4.151742	1.0	3.0	2.0
7	2023-09-29	strategy1	0.160645	32.233188	3.0	1.0	2.0
8	2023-09-29	random	0.174048	4.533993	4.0	2.0	3.0
9	2023-09-29	ew_long	0.160000	-4.247552	2.0	4.0	3.0
10	2023-09-29	strategy2	0.194710	-5.598892	5.0	5.0	5.0
11	2023-09-29	sp500	0.269998	-10.153681	6.0	6.0	6.0
12	2023-10-06	sp500	0.122485	0.281176	2.0	3.0	2.5
13	2023-10-06	gambling	0.059429	-0.969716	1.0	4.0	2.5
14	2023-10-06	random	0.166673	4.199022	6.0	1.0	3.5
15	2023-10-06	strategy1	0.161182	1.873620	5.0	2.0	3.5
16	2023-10-06	strategy2	0.126585	-1.022672	3.0	5.0	4.0
17	2023-10-06	ew_long	0.160000	-1.518810	4.0	6.0	5.0



	group_name	mean_forecasts	mean_decisions	rank_forecasts	overall_decisions	overall_rank
0	sp500	0.136	4.262	2.6	3.1	2.85
1	gambling	0.095	1.044	1.9	4.2	3.05
2	strategy2	0.146	2.074	3.0	3.7	3.35
3	ew_long	0.160	4.263	3.6	3.6	3.60
4	strategy1	0.162	3.473	4.4	3.4	3.90
5	random	0.173	4.048	5.5	3.0	4.25



Overall Performance Across Strategies

Conclusions and Future Endeavors

Conclusions

- Forecasting returns is a challenging work relying only on the market data
- Relying on feature engineering and complex modeling without meaningful justifications may lead to a data snooping issue
- Careful designed portfolio optimization strategy is likely to overperform the benchmarks
- Consistently outperforming all benchmarks over the long term is challenging. To increase the likelihood of achieving better performance, it is necessary to continually evolve both the feature engineering and modeling techniques

Future Endeavors

- Better feature engineering on alpha factors
- Source more alternative datasets, such as, sentiments from news, articles, websites, etc.
- Alternative modeling techniques to better capture linearity and non-linearity relationships
- Alternative Rank forecasting methods
- Alternative decision forecasting methods

Disclaimer

This presentation material is for educational purposes only and does not offer investment advice or pre-packaged trading algorithms. The views expressed herein are not representative of any affiliated organizations or agencies. The main objective is to explore the specific challenges that arise when applying Data Science and Machine Learning techniques to financial data. Such challenges include, but are not limited to, issues like short historical data, non-stationarity, regime changes, and low signal-to-noise ratios, all of which contribute to the difficulty in achieving consistently robust results. The topics covered aim to provide a framework for making more informed investment decisions through a systematic and scientifically-grounded approach.