Predicting Stock Returns based on Convolutional Neural Networks with Feature Operators

Jing-zhi Huang[1] Peilin Liu[2] Jun Ni[3] Jingyi Zhou[4]

[1]Smeal College of Business
Penn State University

[2]College of Engineering
Peking University

[3]School of Economics and Management
Changzhou Institute of Technology

[4]Haıcıoğlu Data Science Institute
University of California, San Diego

CFRI & CIRF Joint Conference
Shanghai
July 30, 2023

Outline

Evidence on ML-based Stock Return Predictability

- Applications of ML
  - Chen and He (2018): CNN can be applied to stock trading data to classify future stock trends
  - Gu et al. (2020) and Leippold et al. (2022): machine learning techniques, such as Random Forest and Neural Network, can be used to combine anomalies
  - Hou et al. (2023): Random Forest and multiple regression outperform the composite score method
- Why ML in China
  - Hou et al. (2023): trading-based signals are typically more effective in the Chinese stock market than accounting-based signals (more suitable for deep learning models)
  - Potential to extract valuable information about trading activity, market microstructure, etc.

Main Question to Be Addressed

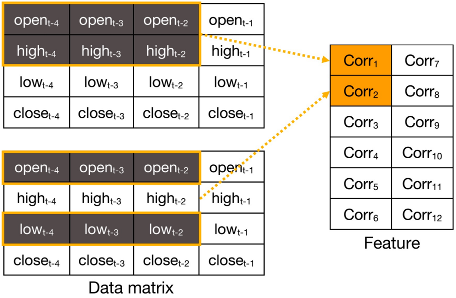
- Factor crowding
  - As the signals are continuously applied to trading, their effectiveness and excess return have been declining and the correlation between new signals and old ones is increasing
- Factor mining directly from raw data
  - Deep learning techniques may be directly applied to stock trading data (avoid manual signal designing)
  - Low signal-to-noise ratio, overfitting, and poor interpretability
- Combining ML with prior financial knowledge (automatic feature engineering)
  - Sharpe (1964): covariance matrix
  - Chan et al. (1996), Bremer and Sweeney (1991), Shimizu et al. (2019), and Huang and Huang (2020): momentum, reversal, low-volatility, moving average, etc.

Focus of Our Study

- A Chinese stock returns prediction problem
  - Independent variable: daily stock trading data in past 30 trading days
  - Dependent variable: 5-day stock returns
- Introducing a CNN model with financial feature operators and constructing better trading-based signals
  - Controlled by previous studies
  - Compared with other ML methods
- Analysis of real market investment
  - Portfolio position optimization

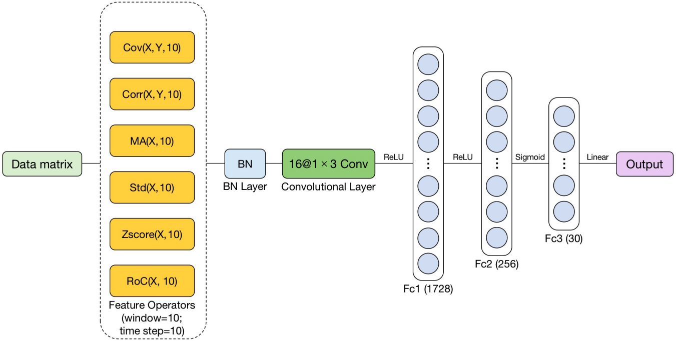
Model Specification

- Feature operators
  - Definition: Cov(X, Y, d), Corr(X, Y, d), MA(X, d), Std(X, d), Zscore(X, d), and RoC(X, d)
  - Mechanism: feature dimension is traversed by permutation approach; time dimension is traversed by time step



Model Specification

- AlphaNet
  - Architecture



  - BN layer: avoid overfitting and increase convergence speed
  - Conv layer: extract higher-dimensional features

Data Used

- Stock trading data
  - Independent variable: open, close, high, low, vwap, volume, return, turnover, free_turnover (9×30 matrix)
  - Dependent variable: 5-day return
- Definition

$$\text{return}_t = \frac{\text{close}_t}{\text{close}_{t-1}} - 1 \qquad \text{turnover}_t = \frac{\text{close}_t \times \text{volume}_t}{\text{total share}_t}$$

$$\text{free\_turnover}_t = \frac{\text{close}_t \times \text{volume}_t}{\text{free float share}_t} \qquad \text{5-day return}_t = \frac{\text{close}_t}{\text{close}_{t+5}} - 1$$

- Time period
  - In-sample period: from Feb. 25, 2015 to Oct. 30, 2018
  - Out-of-sample period: from Oct. 31, 2018 to Nov. 19, 2020

Portfolio Performance

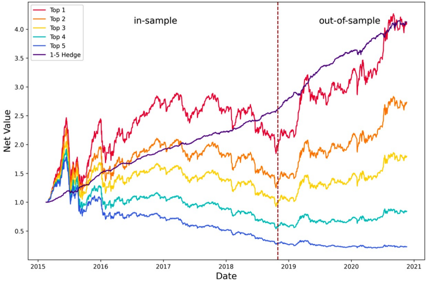
- Equal-weighted portfolio (compounded interest pattern)

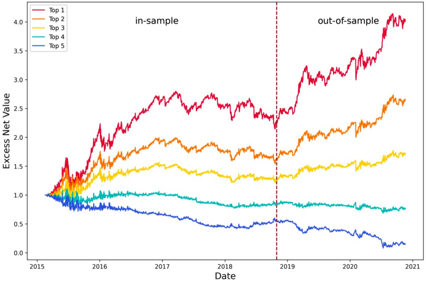
	Annualized return		Sharpe ratio	
	In-sample	Out-of-sample	In-sample	Out-of-sample
Quantile 5	23.30%	38.73%	0.72	1.69
Quantile 4	12.50%	38.05%	0.39	1.64
Quantile 3	4.03%	32.39%	0.13	1.35
Quantile 2	-10.32%	22.22%	-0.32	0.88
Quantile 1	-29.23%	-6.27%	-0.86	-0.23
Long Short	26.30%	22.34%	6.35	4.86

- Stock market is better during the out-of-sample period

Portfolio Performance

- Net value curves





- Good monotonicity (apparent gaps), good generalization (in-sample and out-of-sample), and significant excess return

Control of Risk Model

- Liu et al., 2019: China Four-Factor Model

$$E(R)-R_f = \beta_{MKT} MKT + \beta_{SMB} SMB + \beta_{VMG} VMG + \beta_{PMO} PMO$$

$$R = \alpha + \beta_{MKT} MKT + \beta_{SMB} SMB + \beta_{VMG} VMG + \beta_{PMO} PMO + \varepsilon$$

- OLS regression

	P-value	
	In-sample	Out-of-sample
Intercept	0.000	0.000
MKT	0.355	0.284
SMB	0.002	0.059
VMG	0.003	0.308
PMO	0.582	0.860

Control of China Anomalies

- Hou et al., 2023: 22 trading-based anomaly signals
  - Constructed based only on AlphaNet data to ensure comparability
- Correlation analysis: 11 anomaly signals remained to avoid multicollinearity and inaccurate t-statistics estimation
  - Cross-sectional correlation
  - Time-series correlation

Control of China Anomalies

- OLS regression

	P-value	
	In-sample	Out-of-sample
Intercept	0.000	0.000
size1	0.787	0.134
turn1	0.671	0.529
cvturn1	0.003	0.406
dtv1	0.016	0.014
isc1	0.478	0.020
$isch3_1$	0.460	0.072
ts1	0.079	0.427
cs1	0.395	0.311
betaDM1	0.000	0.009
R1	0.616	0.877
pps1	0.544	0.040

- Significant terms: intercept, dtv1 (liquidity), betaDM1

Models Comparison

- Long-short performance

	Annualized return		Sharpe ratio	
	In-sample	Out-of-sample	In-sample	Out-of-sample
OLS	1.34%	0.95%	0.35	0.31
Random Forest	12.63%	-0.28%	7.79	-0.20
ANN	23.46%	7.66%	2.92	1.46
AlphaNet	26.30%	22.34%	6.35	4.86

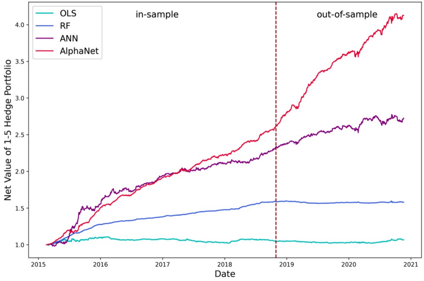
- Long-only performance

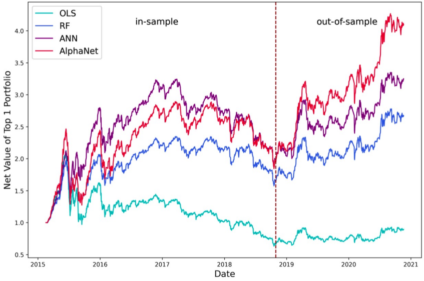
	Annualized return		Sharpe ratio	
	In-sample	Out-of-sample	In-sample	Out-of-sample
OLS	-5.59%	17.62%	-0.17	0.69
Random Forest	16.70%	26.15%	0.66	1.13
ANN	22.30%	26.50%	0.81	1.15
AlphaNet	23.30%	38.73%	0.72	1.69

- Importance of deep learning and AlphaNet's unique design

Models Comparison

- Net value curves





- Better predictor for stock returns and better generalization

Portfolio Position Optimization

- Gaivoronski et al., 2005: Exponential-decay tracking error

$$\mathrm{TE}(w_t) = \sum_{i=t-window}^{t} \left(\sum_{j=1}^{N} w_{t,j} \cdot ret_{t-1,j} - idx\_ret_{t-1}\right)^2 \cdot e^{-\lambda(t-i)}$$

- Optimization task

$$\begin{aligned} & \text{Minimize } \mathrm{TE}(w_t) \\ \text{subject to} \quad & \sum_{i=1}^{N} w_{t,i} = 1 \\ & w_{t,i} = 0 \quad \text{if } alpha_{t,i} < \text{desending}(alpha_t)[300] \\ & 0 \le w_{t,i} \le 0.05 \\ & \frac{\text{sum}(\text{abs}(w_t - w_{t-1}))}{2} \le 0.1 \end{aligned}$$

Portfolio Position Optimization

- Optimization results

	Annualized return	Annualized volatility	Sharpe ratio
Optimal portfolio	51.57%	23.64%	1.88
CSI 500 Index	23.58%	24.75%	0.98
Excess part	22.88%	6.44%	3.24

- Net value curves

Conclusion

- We combine deep learning techniques with prior financial knowledge to achieve prediction for Chinese stock returns
- We test our model under risk models and previous anomalies
- We find our model outperforms other common machine learning methods through its unique design
- We find our model holds the potential to be applied in real market investment to acquire stable excess returns
- These results indicate that our model shows promise as a skillful method to predict stock returns in the cross-section