Empirical Asset Pricing via Machine Learning for Cryptocurrency

2025 Taiwan Risk and Insurance Association (TRIA) Annual Conference

Dr. Li-Han Chang

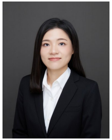
Dept. of Banking and Finance,
Tamkang University

Po-Wei Chen *

Institute of Statistics,
Academia Sinica

Dr. Hung-Wen Cheng

Dept. of Data Science,
Soochow University



Dr. Pei-Jie Hsiao

Dept. of Finance,
National Sun Yat-sen University

Reporter: Po-Wei Chen

December 13 2025

Outline

Introduction: Literature Review

1. Market Efficiency

Cryptocurrency market exhibits signs of inefficiency, implying that it may be possible to design profitable trading strategies based on historical information.

- Urquhart (2016), Nadarajah and Chu (2017), Mensi et al. (2019), Brauneis and Mestel (2018), and Wei (2018).

2. Factor Investing

Cryptocurrency market exhibits several factors that contain valuable information for predicting future returns.

- Hou, Karolyi, and Kho (2011), Asness, Moskowitz, and Pedersen (2013), Liu and Tsyvinski (2021), Bouri et al. (2019), Sakkas and Urquhart (2024), and Sockin and Xiong (2023).

3. Prediction Models

Machine learning has become an effective approach for modeling the complex nonlinear interactions between risk factors and cryptocurrency returns.

- Ibrahim et al. (2021), Baur et al. (2018), Fakhfekh and Jeribi, (2020), Chen et al. (2021), Khedr et al. (2021), Gu, Kelly, & Xiu (2020), Greaves and Au (2015), Indera et al. (2017), Lee (2017), Liu et al. (2021), Xiaolei, Mingxi, and Zeqian (2020), and Chen et al. (2021).

4. Information Coefficient (IC)

IC serves as a measure of forecasting ability of market indicators or as a practical metric for model effectiveness in financial prediction.

- Ambachtsheer and Farrell (1979), Ding (2011), Ding and Martin (2017), Zhang and Lu (2024) , and Ding et al. (2024).

2. Data and Factors

- Data
- Factors

Data and Factors: Data and Sample Splitting

Data Source: CoinMarketCap.com

Our Sample Selection Criteria:

- Top 1000 coins by market capitalization.
- Have nonzero market value, trading volume, and price.

Finally 998 coins are left. And we extract the daily Open, High, Low, Close, Volume, and Market Capitalization data.

Sample Splitting

- Our sample covers the period from June 2018 to March 2022. Total 197 weeks.
- Divide the 197 weeks into 97 weeks of training sample and the remaining 100 weeks for out-of-sample testing.

Data and Factors: Factors

- We choose the factors selected in Liu, Tsyvinski and Wu (2022) that generate statistically significant returns from long-short strategies.

Category	Factor	Definition
Size	MACP	Log last-day market capitalization in the portfolio formation week
Size	PRC	Log last-day price in the portfolio formation week
Size	MAXDPRC	Maximum price of the portfolio formation week
Momentum	r 1,0	Past one-week return
Momentum	r 2,0	Past two-week return
Momentum	r 3,0	Past three-week return
Momentum	r 4,0	Past four-week return
Momentum	r 4,1	Past one-to-four-week return
Volume	PRCVOL	Log average daily volume times price scaled by market capitalization in the portfolio formation
Volatility	STDPRCVOL	Log standard deviation of price volume in the portfolio formation week

3. Methodology

- Information Coefficient (IC)
- Machine Learning Models
- Investment Strategy Framework

Methodology: Information Coefficient (IC)

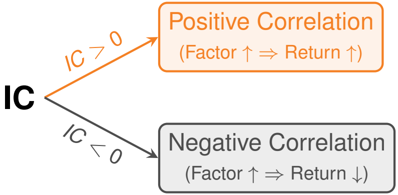
The IC is defined as

$$IC_{k,t+h} = \text{Corr}(\mathbf{X}_{k,t}, \mathbf{R}_{t+h}),$$

where

- $X_{k,t}$ denote the value of factor k at time t.
- R_{t+h} denote the realized return over the holding period from t to $t+h$.
- $\text{Corr}(\cdot)$ represent the correlation operator.

The IC measures both the direction and the magnitude of the linear relationship between a factor and subsequent returns.



Methodology: Machine Learning Models

In its most general form, we describe a factor's IC as an additive prediction error model:

$$IC_{k,t+h} = \mathbb{E}[IC_{k,t+h}] + \epsilon_{k,t+h},$$

where the conditional expectation is modeled as:

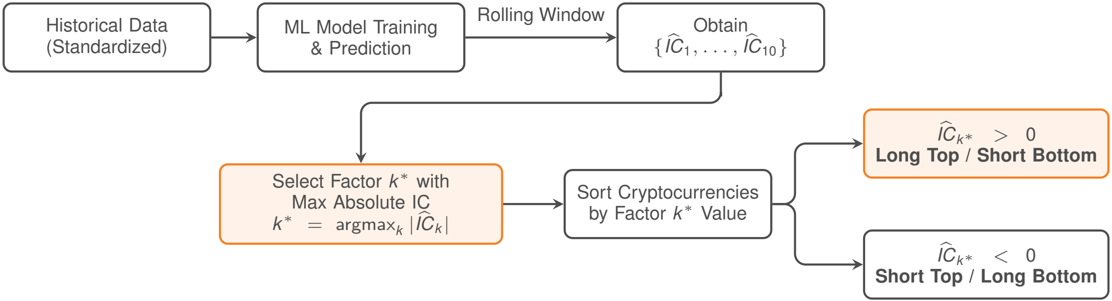
$$\mathbb{E}[IC_{k,t+h}] = g^*(z_{k,t})$$

- $z_{k,t}$ denotes the vector of predictors (factor values).
- $g^*(\cdot)$ represents the linear or nonlinear predictive function approximated by ML models.

List of Candidate Models

Linear	**Tree-Based**	**Kernel / Instance**	**Deep Learning**
OLS	RF, DT GBDT, LGBM	SVR / KNN	Neural Networks (1–5 Layers)

Methodology: Investment Strategy Framework



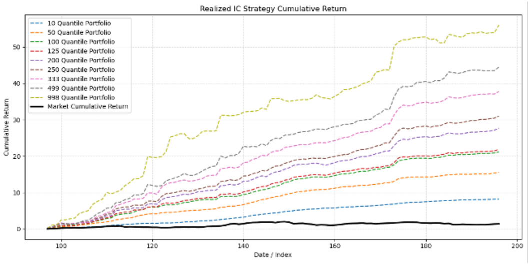
Note: We construct quantile portfolios (e.g., 10, 50, 100, 125, 200, 250, 333, 499, 998 groups) based on the sorted factor values.

4. Empirical Results

- Realized IC-Based Strategy
- Predicted IC-Based Strategy

Empirical Results: Realized IC-Based Strategy

- The win rates of all quantile portfolios exceed 70%, with several surpassing 80%.
- The cumulative returns increase monotonically as the number of quantiles grows.
- If we can accurately predict the magnitude and direction of the ICs for each factor, we can generate meaningful cumulative returns.



Quantile Portfolio	Win Rate	Cumulative Return
10 quantile portfolios	71.52	8.2294
50 quantile portfolios	81.24	15.5937
100 quantile portfolios	81.48	21.2015
125 quantile portfolios	79.27	21.8907
200 quantile portfolios	78.86	27.7005
250 quantile portfolios	80.13	31.0385
333 quantile portfolios	80.52	37.7998
449 quantile portfolios	75.70	44.5102
998 quantile portfolios	71.58	56.0030
Equal-weighted Market Portfolio	–	1.4224

Empirical Results: Predicted IC-Based Strategy

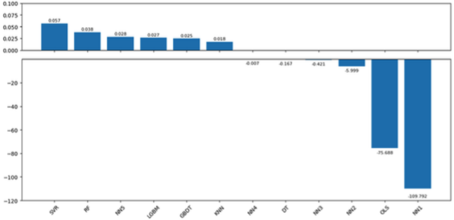
Evaluation Metric:

$$R^2_{\text{oos}} = 1 - \frac{\sum_{t \in \mathcal{T}} \sum_{k=1}^{K} \left(\text{IC}_{t,k} - \widehat{\text{IC}}_{t,k} \right)^2}{\sum_{t \in \mathcal{T}} \sum_{k=1}^{K} \text{IC}_{t,k}^2},$$

where $\mathcal{T}$ denotes the testing sample and $K = 10$ represents the number of factors.

- SVR and tree-based methods achieve superior prediction performance.
- As the number of layers in the neural networks increases, the R^2_{oos} consistently improves.

	OLS	RF	DT	GBRT	LGBM	SVR	KNN	NN1	NN2	NN3	NN4	NN5
All	-75.67	0.038	-0.17	0.025	0.027	0.057	0.018	-109.8	-5.999	-0.421	-0.007	0.028

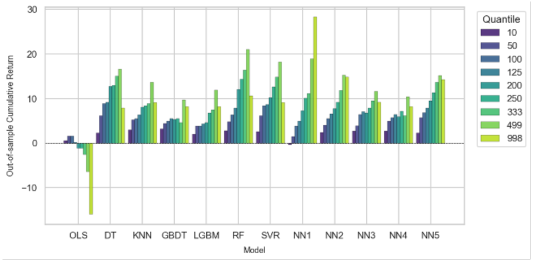


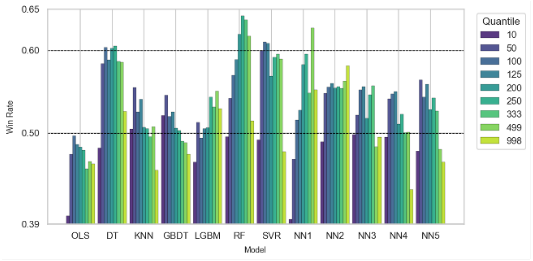
Out-of-sample R^2 across different machine learning methods.

Empirical Results: Predicted IC-Based Strategy Cont.

Cumulative Return and Win Rate

- In general, all models exhibit a clear upward trend in cumulative returns as the number of quantiles increases.
- RF, DT, and SVR present the strongest overall performance among all models.
- NN1 achieves good returns in the 499- and 998-quantile portfolios but performs poorly in the other settings.
- As the number of layers in the neural networks increases, the results become more stable and robust across all quantile portfolios.





Empirical Results: Comparison between Predicted and Realized

- Although RF, DT, and SVR present strong results in our strategy framework.
- A noticeable gap remains between the predicted and the realized IC results, suggesting that there is still room for improvement when applying this strategy.

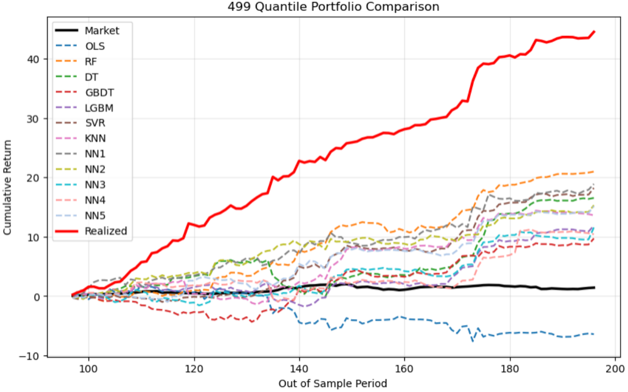


Figure: 499 Quantile Portfolio Comparison

Conclusion

- This study develops a data-driven framework for forecasting information coefficients (ICs) in the cryptocurrency market and constructing IC-based long-short portfolios.
- We show that RF, DT, and SVR deliver superior return and win rate under this framework.
- Finer quantile portfolios achieve higher cumulative returns and win rates, whereas coarser portfolios perform less effectively.

Thank You!

Po-Wei Chen
brianchen1229@gmail.com