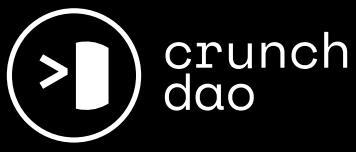


CROWDSOURCED INVESTMENT RESEARCH

BY CRUNCHDAO

Matteo Manzi, Enzo Caceres, Correlator, Kain, SRK,

fortunefavorsthebrave

2022/11/01

CROWDSOURCED INVESTMENT RESEARCH

abstract

Markets are complex, high-dimensional, chaotic, stochastic, non-Gaussian dynamical systems. With a data-driven perspective, CrunchDAO's tokenomics powers a crowdsourced investment strategy that, via Machine and Ensemble Learning, leads to competitive financial services.

Crowdsourced Investment Research

Introduction

In the context of quantitative finance, in which a systematic, rationalized, quantitative (i.e., scientific) approach is preferred over discretionary decision-making, the field of machine learning is getting a lot of traction (Prado 2018). This is because of its natural compatibility with a Bayesian approach (Barber 2012), characterizing the field of econometrics, but also because of its potential to construct nonlinear models from financial data (Chan and Mátyás 2022).

In this context, CrunchDAO proposes an Ensemble Learning framework via tournaments ((Craib et al. 2017), (Prado, Prado, and Fabozzi 2019)), in order to generate *market-neutral* signals.

In this section, we should make the case of ensemble learning in quantitative finance. Particularly discussing how, in the context of ensemble learning and bagging in particular, combining a variety of orthogonal models yields more accurate estimates of expectations.

Traditional Econometrics and Risk

$$r_i = \sum_k X_{ik} f_k + u_i \tag{1}$$

Data

CrunchDAO makes use of different datasets.

- C-MECHANICS: This strategy is a trend-following strategy based on the trend of idiosyncratic (an individualizing characteristic) return and volatility.
- E-KINETIC: This momentum outlook aims to systematically isolate and harvest excess returns arising from behavioral market anomalies by investing in diversification, not performance.

- B-VOLATILITY: This strategy identifies distortions in volume, price, and volatility between short-dated options and stock prices.
- 3B1-SIGNAL: Institutional investors are leveraging equity factor risk models (Sector / Country Stock etc.) to predict return and hedge their bets. We investigate the extent to which nonlinearities not captured by standard linear models within equity factor risk models are present. Some generated factor returns and information ratios higher than corresponding linear factors
- DOLLY: Portfolio managers invest a tremendous amount of time and resources in identifying equity that will outperform the market in the long term - alpha- ; In Dolly, the community leverages machine learning to select top long-term asset managers and piggyback their trades. Securities and Exchange Commission (SEC) 13f filing data offer valuable insight into top asset managers' holdings at each quarterly filing point.
- GORDON-GEEKO: This strategy uses trade information from top management and senior executives (i.e. insiders) as it has been demonstrated in past academic research that insiders have insight - or alpha - over other investors.

Feature Engineering

CrunchDAO's Machine-Learning-enabled ensemble framework builds on top of traditional econometric risk models, requiring a number of steps in the data preparation: features orthogonalization, standardization, model order reduction and data obfuscation will be discussed.

Data Obfuscation

Data anonymization is performed using quantization schemes.

Here is a non-esaustive list of interesting projects we have been researching and that could provide interesting tools to challange the current design choices:

- PySyft
- Weavechain
- BeekeperAI
- Microsoft SEAL and associated compiler
- ZKML
- Zama

Tournament

Staking Model

In this section we introduce two objectives, a performance measure that could determine how good a prediction is and a diversity measure giving us a degree of orthogonality among predictions. The combination of these two requirements inform the ensemble learning step, following in the metamodel pipeline.

One of the tournament's main constraint is keeping the reward system robust to Sybil attacks. One way to do it is to make having multiple accounts costly and futile. In other words to have each participant have skin in the game. In this scenario, the participant needs to lock some $CRUNCH in a smart contract, leading to a reward proportional to the amount of his/her stake.

1. Inspired from Validator-Delegator model of Cosmos blockchain.
2. In CrunchDAO context: Validators -> MM heros, Delegators -> MM supporters
3. 100 Heros with the highest total stake are chosen for creating the stakeweighted MM.
4. All holders of the Crunch token can act as supporters by backing their fav hero. Supporters are eligible for getting a cut of payouts from the hero pool they staked on. (Hero stake + supporter stake considered for top 100 selection)
5. To avoid frequent stake switches by supporters, their stake is locked for 1 month.
6. Heros can decide on their commission and a fixed fees. This will allow for competition between heros to attract supporters (and be among the top 100 staked).
7. Hierarchical Clustering can be used to find submissions belonging to a common cluster. Submissions belonging to the same cluster should be penalised.
8. As tournament progresses and it gets more unique submissions, the limit of 100 heros can be increased through community vote.

A set of proposal are currently being discussed to find a solution for Sybil attack resistance: (Li et al. 2017).

Why this scheme makes sense:

1. Encourages competition among heros and motivates them to continuously improve their models to attract supporters.
2. More involvement from general token holders who believe in the project, not just modellers.
3. Token holders will want the MM to do well, hence will stake on the model which contributes positively.
4. Good for overall tokenomics. As most of the tokens will be staked.
5. Payouts create positive feedback and can be restaked instead of selling in open market.
6. Unique scheme not followed by any of the competitors.

Fineprints:

1. Burning should be enabled post staking.
2. Stake cap on models. If stake limit is reached, supporters can't stake on that model, they have to stake on some other model. This will encourage decentralization and diversification.
3. Payout factor decreases if the model is part of an existing cluster. i.e if 5 models fall in the same cluster, payout factor=1/5 for all the 5 models. This would also encourage supporters to stake somewhere else i.e. stake on unique models.

Sybil attacks will be avoided due to staking. And supporters will tend to stake on a reliable model which will further prevent sybil attacks. Payouts will only be made to staked (possibly top 100 staked) models, so no value in pursuing sybil attacks.

Problems with K-Means clustering in CrunchDao context:

1. K-means clustering is dependent on how it selects the initial points as cluster centroids. So, entirely different clusters are possible just by changing the random seed. Similar submissions might fall in different clusters in one run, and in same cluster in another run. This unpredictability is undesirable (and might seem unfair if unexpected results show up).
2. K-Means requires the 'K' parameter, i.e. we need to specify upfront how many clusters should be formed. This is non-trivial and requires another level of analysis called the 'Elbow method'; this is a visual analysis for finding the appropriate value of 'K'. This again is a subjective choice and somebody will be required to select that value for each round, this again is undesirable.

Benefits of Hierarchical clustering (Agglomerative clustering) in CrunchDao context:

1. No randomness. It starts by forming clusters of the most similar submissions and then forms bigger clusters from these small clusters in a hierarchical way. This will always produce the same clustering results.
2. No need to specify any hyperparameter. A threshold needs to be applied at the end of the process to form clusters. In our context, let's say we want all submissions which are more than 90% correlated to fall in the same cluster, then we put the threshold of 'correlation distance' as (1 - 0.9). This is easy to explain and interpret.

Example:

- Model1
  - Total Stake: 1000
  - Part of Cluster 1
  - Performance metric (spearman correlation): 0.03

- Model 2
  - Total Stake: 800
  - Part of Cluster 2
  - Performance metric (spearman correlation): 0.02
- Model 3
  - Total Stake: 500
  - Part of Cluster 1
  - Performance metric (spearman correlation): 0.03
- Payout multipliers
  - Model 1, Model3 : 0.5 (Since they belong to same cluster)
  - Model 2: 1.0 (Unique cluster)
- Payouts: (Stake * Performance * Multiplier)
  - Model 1: 1000 * 0.03 * 0.5 = 15
  - Model 2: 800 * 0.02 * 1.0 = 16
  - Model 3: 500 * 0.03 * 0.05 = 7.5
- Payout distribution to stakers (Model 2 example)
  - Total payouts for Model 2 = 16
  - Num Supporters = 3 (say)
  - Num Heroes = 1 (always)
  - Hero Stake = 50 ; Supporters stake: 250 * 3 = 750 (assuming each supporter stakes 250 each)
  - Hero commission (irrespective of Hero stake) = 25% (to be chosen by Hero)
  - Commission going to Hero upfront = 16 * 0.25 = 4
  - Remaining Payouts : 16 - 4 = 12
  - 12 is distributed to all Supporters + Hero based on their stake
  - Hero gets : $\frac{50}{800}$ * 12
  - Each supporter gets: $\frac{250}{800}$ * 12
  - Hero commision is in addition to Hero stake payouts. Both are necesssary beacuse Hero commission will lead to competition among Heros to provide best service at lowest commission to attract supporters; and Stake payouts are necessary as they will indicate to the supporters that the Hero has skin in the game and is willing to lose money if performs badly

Note: Spearman Correlation is a metric that can be used as a starting point. The performance metric will need to be evaluated periodically whether it benefits the Fund performance.

Query on discord: Are both Fineprints 2 and 3 needed?

- Short Answer, Yes.
- Fineprint 3 is needed to disincentivise modellers submitting similar predictions.
- Fineprint 2 is needed to avoid over allocation of capital in a few models. Say over the last month a model gives outsized returns with high originality, then supporters may unsuspectingly want to switch their stakes on this model (which might be highly volatile and detrimental for future rounds). This will unwantedly lead to over allocation in a single model. Hence, having a stake cap on models will be useful. The amount of stakecap can be reviewed from time to time.

Alpha provider scheme is more nuanced because:

1. Providers may not want to give away their new feature for everyone to use.
2. The feature might become less useful overtime due to alpha decay, and Crunch team will have to deal with the evaluation of such decayed features and remove it from the dataset. This will lead to more manual evaluation from the Crunch team side which is a bottleneck for improvement.

Having said that, it can be taken up as 2 step process:

1. Evaluation of utility by the provider. Crunch team provides an api which can be used by providers to upload their feature (with stock tickers), the api returns the correlation of the queried feature with the existing features in the DataCrunch dataset. (not the correlation with target). This will let the providers know if their feature is unique or not.
2. Say a provider develops a new feature which is unique wrt all other features in the dataset. Then, the provider can request a custom dataset (obfuscated like existing datasets) which includes that new feature. The provider is then free to use the new feature however they like to make predictions as usual for the tournament. For requesting a custom dataset, a small amount of Crunch can be locked (not staked) to avoid spamming.

Originality is directly included in the reward computation as discussed. Since the payout is only for the staked users, using multiple accounts will be discouraged as there is no incentive in submitting from multiple accounts as the originality factor will reduce if ones does that.

Multi-objective Optimization Problem and Gamification The need to both optimize for performance and originality naturally leads to the definition of a multi-objective optimization problem, for which the Pareto Frontier can be computed.

A Pareto optimal submission is a non-dominated vector in the vector space of all feasible submissions. These submissions are associated with a rewarding factor $\alpha > 1$, boosting the reward of submissions

on the frontier associated with positive performance. tive.

The Scoring System

The current Scoring metric is the Spearman's rank correlation coefficient between the submission and the realized targets.

Metamodeling

Clustering and Dimensionality Reduction

(Avellaneda 2019), (Akansu, Avellaneda, and Xiong 2021) will be discussed here.

Ensemble learning

A straightforward approach is to compute a weighted average of all models based on their stakes. In mathematical terms, let's say we have n participants and participant u (P_u) has staked C_u Crunch tokens where u=1, 2, …, n so the amount of involvement of each participant in the metamodel is

$$w_u = \frac{C_u}{\sum_{u=1}^{n} C_u}$$

The final metamodel would be the weighted combination of the predictions submitted by participants. In this case. If S_u represents prediction of P_u then final prediction (S) is

$$S = \sum_{u=1}^{n} w_u S_u$$

Moreover, the statistics of the set of predictions can be used to infer a measure of risk in the portfolio management process. We discuss how to integrate this in modern portfolio theory. We briefly discuss the necessary relation between these design choices and the ergodic hypothesis on financial.

Discuss Unscented Transform and its relation to being able to estimate the portfolio risk in a variance sense using nonlinear models, like in the Unscented Kalman Filter and Particle methods (Blackmore 2006).

Same analogy as using high order surrogate models to propagate uncertainties even if just interested in mean and covariance (Vasile and Manzi 2022), (Manzi and Vasile 2020).

Porfolio Optimization

(Stuart and Markowitz 1959), (Chriss and Almgren 2005), (Crama and Schyns 2003), (Pafka et al. 2004), (Lobo, Fazel, and Boyd 2007), (Acikmese, Carson, and Blackmore 2013)

References

Acikmese, Behcet, John M. Carson, and Lars Blackmore. 2013. "Lossless Convexification of Nonconvex Control Bound and Pointing Constraints of the Soft Landing Optimal Control Problem." *IEEE Transactions on Control Systems and Technology*. https://doi.org/10.1109/tcst.2012.2237346.

Akansu, Ali N., Marco Avellaneda, and Anqi Xiong. 2021. "Quant Investing in Cluster Portfolios." *Journal of Interaction Science*. https://doi.org/10.21314/jois.2021.006.

Avellaneda, Marco. 2019. "Hierarchical Pca and Applications to Portfolio Management." *Revista Mexicana de Economía Y Finanzas*. https://doi.org/10.21919/remef.v15i1.446.

Barber, David. 2012. "Bayesian Reasoning and Machine Learning." https://doi.org/10.1017/cbo9780511804779.

Blackmore, Lars. 2006. "A Probabilistic Particle Control Approach to Optimal, Robust Predictive Control." https://doi.org/10.2514/6.2006-6240.

Chan, Felix, and László Mátyás. 2022. "Econometrics with Machine Learning." *Advanced Studies in Theoretical and Applied Econometrics*. https://doi.org/10.1007/978-3-031-15149-1.

Chriss, Neil, and Robert Almgren. 2005. "Portfolios from Sorts." https://doi.org/10.2139/ssrn.720041.

Craib, Richard, Rey Bradway, Xander Dunn, and Joseph Krug. 2017. "Numeraire: A Cryptographic Token for Coordinating Machine Intelligence and Preventing Overfitting."

Crama, Yves, and Michael Schyns. 2003. "Simulated Annealing for Complex Portfolio Selection Problems." *European Journal of Operational Research*. https://doi.org/10.1016/s0377-2217(02)00784-1.

Li, Ming, Ming Li, Ming Li, Jian Weng, Anjia Yang, and Wei Lu. 2017. "CrowdBC: A Blockchain-Based Decentralized Framework for Crowdsourcing." *IACR Cryptology ePrint Archive*. https://doi.org/10.1109/

tpds.2018.2881735.

Lobo, Miguel Sousa, Maryam Fazel, and Stephen Boyd. 2007. “Portfolio Optimization with Linear and Fixed Transaction Costs.” *Annals of Operations Research*. https://doi.org/10.1007/s10479-006-0145-1.

Manzi, Matteo, and Massimiliano Vasile. 2020. “Analysis of Stochastic Nearly-Integrable Dynamical Systems Using Polynomial Chaos Expansions.” In.

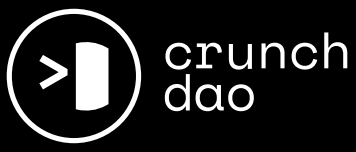
Pafka, Szilard, Marc Potters, Marc Potters, and Imre Kondor. 2004. “Exponential Weighting and Random-Matrix-Theory-Based Filtering of Financial Covariance Matrices for Portfolio Optimization.” *arXiv: Statistical Mechanics*.

Prado, Marcos Lopez de. 2018. “Advances in Financial Machine Learning.”

Prado, Marcos Lopez de, Marcos Lopez de Prado, and Frank J. Fabozzi. 2019. “Crowdsourced Investment Research Through Tournaments.” *The Journal of Financial Data Science*. https://doi.org/10.2139/ssrn.3454234.

Stuart, Alan, and Harry M. Markowitz. 1959. “Portfolio Selection: Efficient Diversification of Investments.” *A Quarterly Journal of Operations Research*. https://doi.org/10.2307/3006625.

Vasile, Massimiliano, and Matteo Manzi. 2022. “Polynomial Stochastic Dynamical Indicators.” *Celestial Mechanics and Dynamical Astronomy*, December.


crunch
dao