Oracle Counterpoint:

Relationships between On-chain and Off-chain Market Data

Zhimeng Yang, **Ariah Klages-Mundt**, Lewis Gudgeon

Coinbase Cornell University, Imperial College London

MARBLE, July 2023

The oracle problem

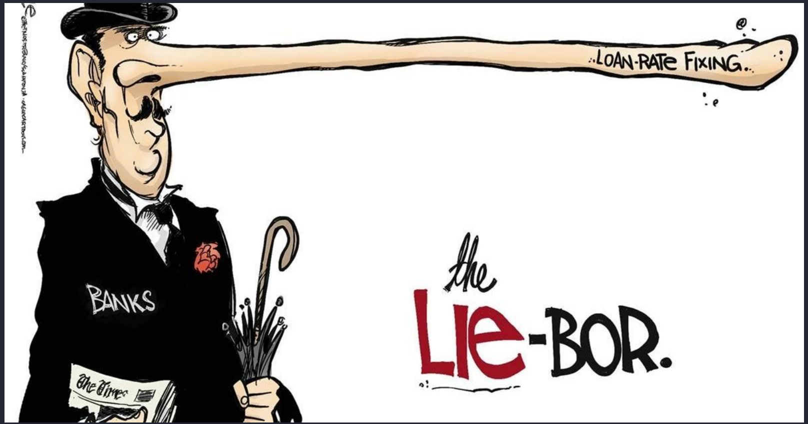
- Financial contracts require secure data feeds
- How do we build these in resilient ways?
- Not unique to blockchain

Example: LIBOR manipulation (2003-2012)

- Manipulate interest rate data used in many contracts
- Deutsche Bank, Barclays, Citi, JP Morgan +
- Banks submitted incorrect data to force LIBOR to their advantage (make positions profitable)

Takeaways

- Known issue even with high reputation entities
- Can blockchain help us build more resilient feeds?

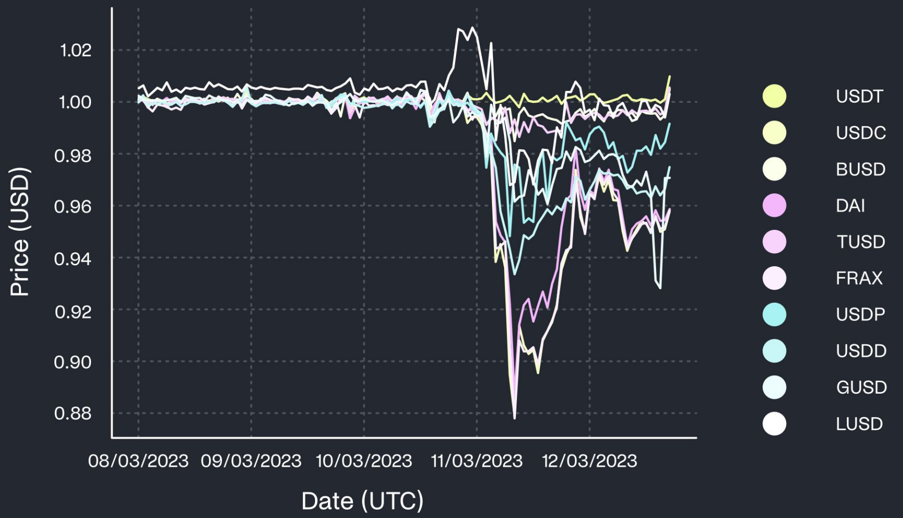


In blockchain applications: need price data

- Some price processes take place off-chain (on CEXs)
- And these processes involve off-chain assets (USD) and *can't* happen on-chain
- Important: dangerous to use stablecoins as quote asset in place of USD (example: Mar 2023)

Need oracles to import these prices on-chain

- Oracle price correctness can't be fully verified
- Can only authenticate that provider is who they say they are



All of DeFi relies on oracles

Stablecoins
Price their reserve assets

Insurance
Trigger condition and how much to pay out

Lending
Compute collateralization, trigger liquidation

Derivatives
Compute payments, trigger position closure

...

Current oracle approaches come with challenges

Centralized Oracles
Requires trust in a central data provider.

Medianizing (Chainlink)
Off-chain aggregation + on-chain verification. →
Essentially a trusted multisig, potential collusion.

DEX TWAPS
- Manipulable when liquidity is low (e.g. INV, Apr 2022), more so after the merge, but quantifiable costs
- slow
- only for on-chain assets (no USD)

Betting Markets / Data Derivatives
Potential collusion / Keynesian beauty contest

Topics

I. Motivation: validating *likelihood* of oracle integrity

II. Research into new data sources for such approaches

III. Future directions

New approach: validating *likelihood* of oracle integrity given other observable variables

Input data

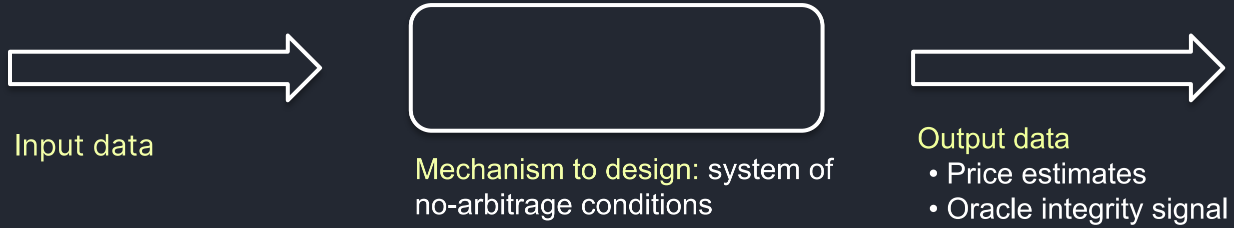
- On- and off-chain sources
- Price and possibly non-price data

Mechanism that we need to design: system of no-arbitrage conditions on input data

Output data

- Price estimates
- Signal about likelihood of oracle integrity

Motivation / use case of this design



Treat oracle data as *candidate* prices and consolidate with on-chain data.
Goal: *fast* but with added safety guarantees!

Protocols can use oracle integrity signal (new info) to improve security:

- Trade off liveness ↔ economic security
  - ⚠ Current Default: Security ?, Liveness ↑↑
- Decide which oracle they want to use

Input data

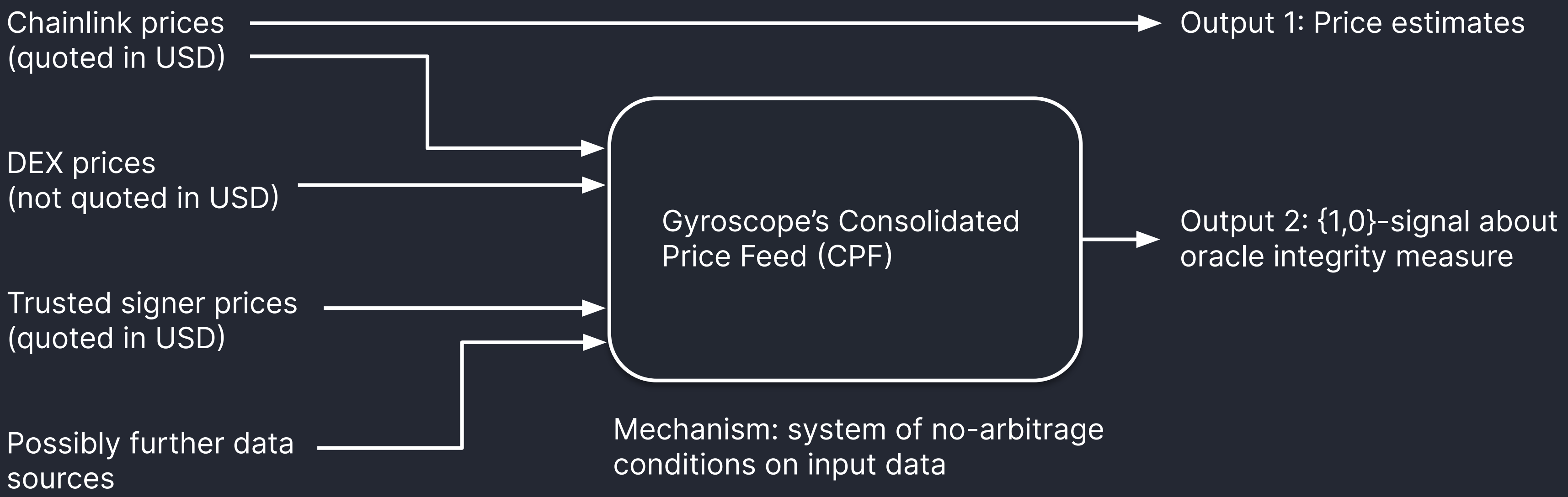
Mechanism to design: system of no-arbitrage conditions

Output data
- Price estimates
- Oracle integrity signal

What this mechanism should aim to do:
- Speed
- Liveness
- Cost to manipulate price estimates
- Cost to manipulate signal of oracle integrity
    - A DoS: potentially affect liveness of protocols using the price feed

Why this is hard:
- Formulating system of no-arbitrage conditions to get these properties
- Balance security models, manipulation costs, failure points of different data sources
- Cover corner cases of stablecoin pricing

First version of this new approach: Gyroscope's Consolidated Price Feed (CPF)

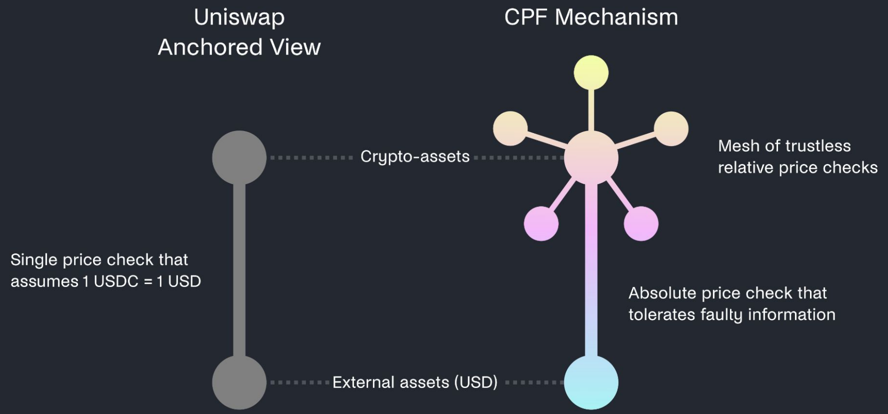


Vector of Oracle Prices = a solution to a system of equations

Oracle USD prices p_i solve

$p_{i/j}$, some on-chain source $\approx \dfrac{p_i}{p_j}$ relative

$\tilde{p}_i$, some on-chain source $\approx p_i$ absolute



For at least a spanning tree of (i,j) pairs.
Pull $p_{i/j}$ from a resilient, but not perfect, data source.
No arbitrage bounds.
Can use multiple checks.

For at least one i.
Pull $\tilde{p}_i$, some source from a resilient, but not perfect, data source.
Can use multiple checks.

Our motivation: Absolute Price Checks

- Choose reference asset i.
  - ⚠ Absolute price sources are significantly error/manipulation-prone!
  - ➔ Use multiple!
- DEX TWAPS: i/b pairs where b is a stablecoin.
- Signed (centralized) prices $p_{i,signed}^{k}$

Question: Are more on-chain sources possible?

II. Research into new data sources to augment this style of oracle

[Submitted on 28 Mar 2023]

Oracle Counterpoint: Relationships between On-chain and Off-chain Market Data

Zhimeng Yang, Ariah Klages-Mundt, Lewis Gudgeon

https://arxiv.org/abs/2303.16331

Can non-price chain data help sense-check prices?

- Aim: recover off-chain price signal from non-price chain data
- Context: agents incorporate off-chain prices (e.g., ETH/USD) into on-chain decisions
    - *Some* causal relationship here likely (both in economic models and empirical)
- Try to recover this price information by analyzing on-chain activity
- Want: alternative trustless input for absolute price checks
- This data can be manipulated, but costly to do so (part of on-chain markets)

Basic on-chain data:

- Block and tx data
- ETH circulation measures
- Network computational consumption (gas market)

DEX participation measures (non-price)

Transformed features informed by economic models

Function *f*

Recovered ETH/USD information

Goal is to find a good function *f* to do this

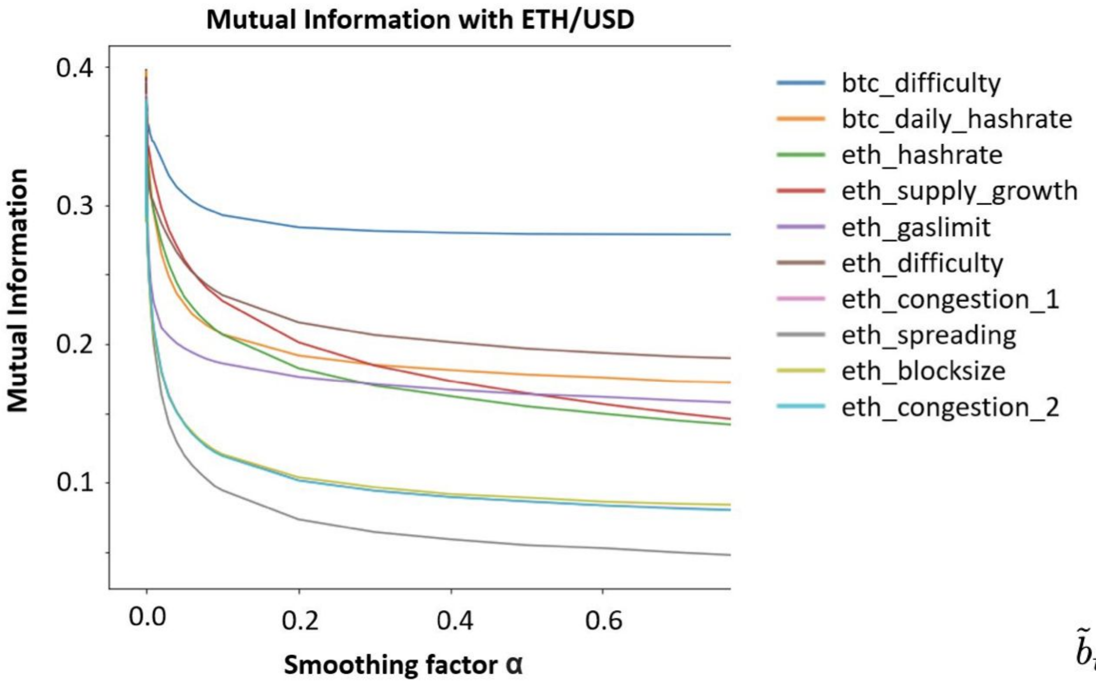
Data Sources

	CCY	Source	Starting From	Frequency
On-Chain	BTC block data	Google BigQuery	Jan 2016	Hourly
	ETH block data			
	CELO block data	Celo Graph (block, celoTransfers)	Apr 2020	
	cGLD transaction data			
	cUSD transaction data			
	Uniswap liquidity and balance data	The Graph	Aug 2020	
Off-Chain	BTC price and volume data	Coinbase API	Jan 2016	
	ETH price and volume data			
	Celo price and volume data		Sep 2020	

Exploring relationships between chain features and ETH/USD price (both at same time *t*)

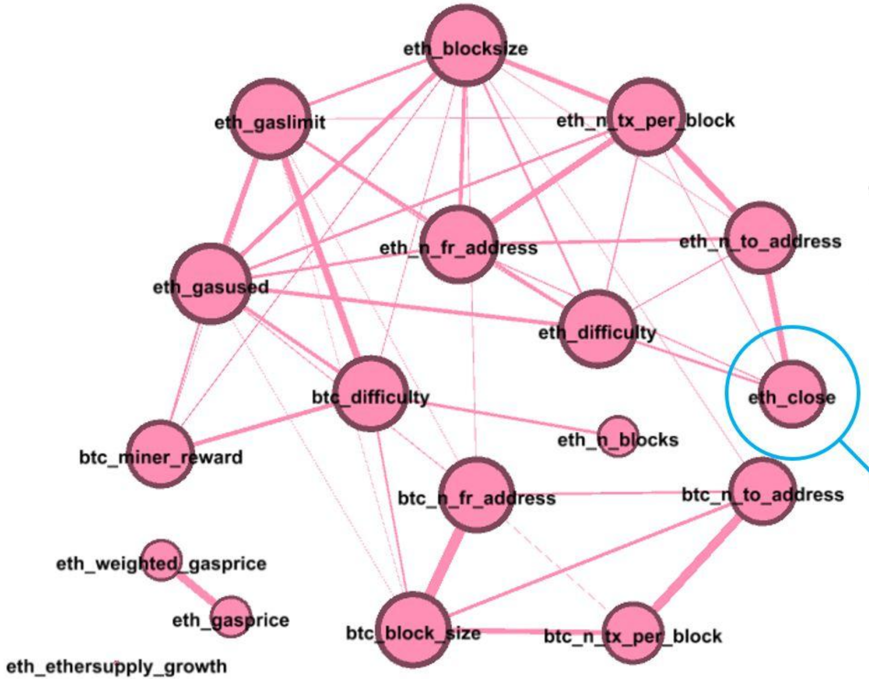
- Mutual information = reduction in uncertainty (info obtained) about X by observing Y
  - Entropy measures how surprising typical outcome of a variable is = information value

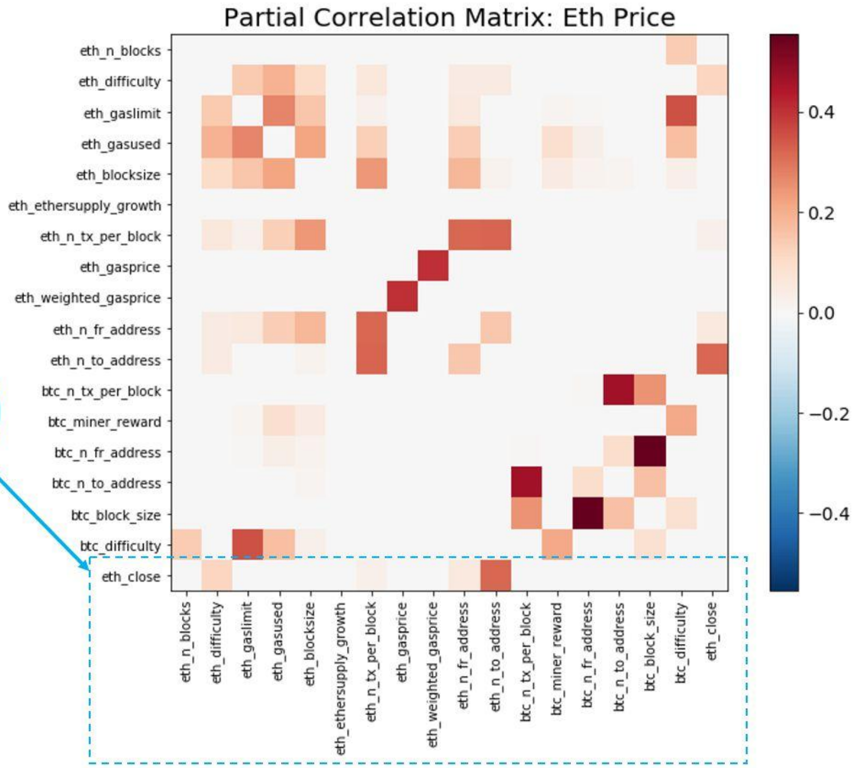
- Sparse Inverse Covariance Estimation
  - Probabilistic model of partial/pair-wise relations between variables
  - If true underlying structure is Gaussian, entries of inverse covariance matrix are zero iff variables are conditionally independent
  - If not Gaussian, then just get partial correlations



$$\tilde{b}_t = (1-\alpha)b_t + \alpha\tilde{b}_{t-1}.$$

- Several variables appear to contain info relevant to price
- Smoothed data generally less informative than the most up-to-date data (perhaps intuitive)

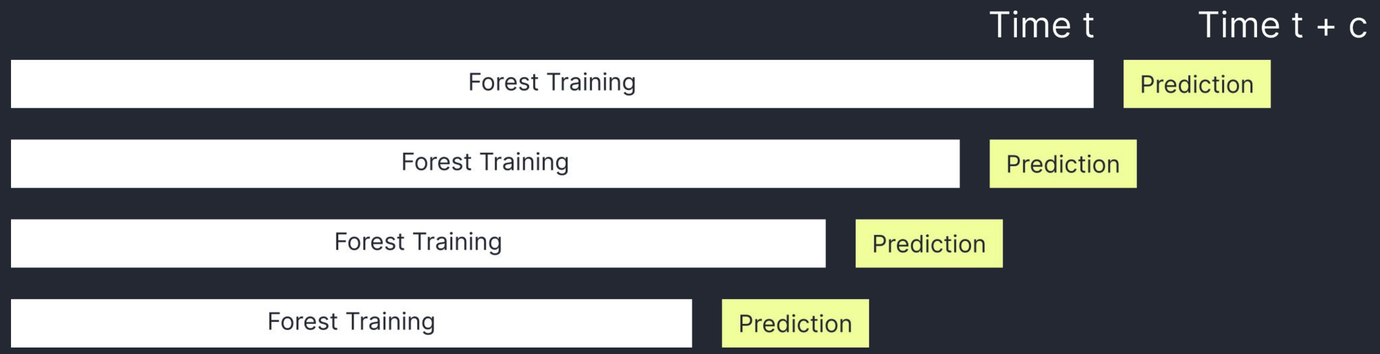




- Some strong partial correlations with price
- Others may be indirectly related via effects on other variables (if the graphical model is correct)

Modeling ETH/USD price from on-chain data

- Tree ensemble methods on rolling training-testing data split
- Not a prediction of future prices, but try to recover signal of *current* price given *current* chain features



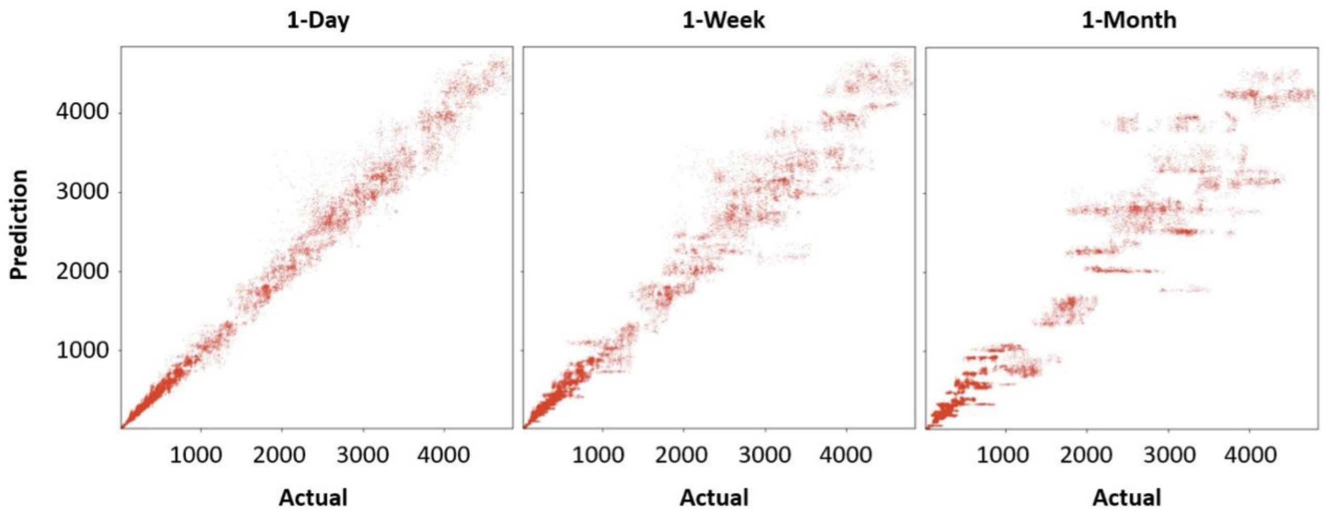


Fig. 5: Recovered price vs actual for random forest with given retraining periods.

- Noisy signal, some price information recovered
- 1-day looks best, but 1-month better performance by some measures

Measuring performance of on-chain price recovery

- Compare against martingale benchmark (efficient market)
  - Suppose last observed price in last retraining period is best estimate of next price, barring new info
  - Consider on-chain data as only source of new info

- Measuring squared error vs true prices:

  $SE = (predicted/actual - 1)^2$

- Mean squared error (MSE) over different sets of time t
  - From time t_s when there is sufficient training data
  - Further restricting to top 10% volatility times

- Difference in squared errors:

  $\mathrm{DSE} = (benchmark/actual - 1)^2 - (model/actual - 1)^2$

Model retraining periods:	1-day	7-day	30-day
How often model beats benchmark	12.4%	26.9%	32.4%
Gain over benchmark when model is better	0.65%	3.56%	7.10%

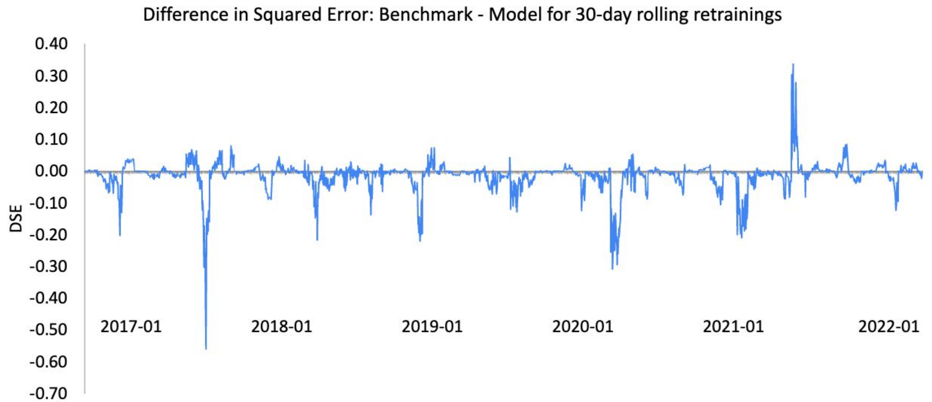
Table 1: Summary of DSEs between models and benchmarks for different retraining periods evaluated on the whole dataset (2016-2022). Row 1 is the frequency that $DSE > 0$. Row 2 is the root mean DSE at the times that $DSE > 0$.

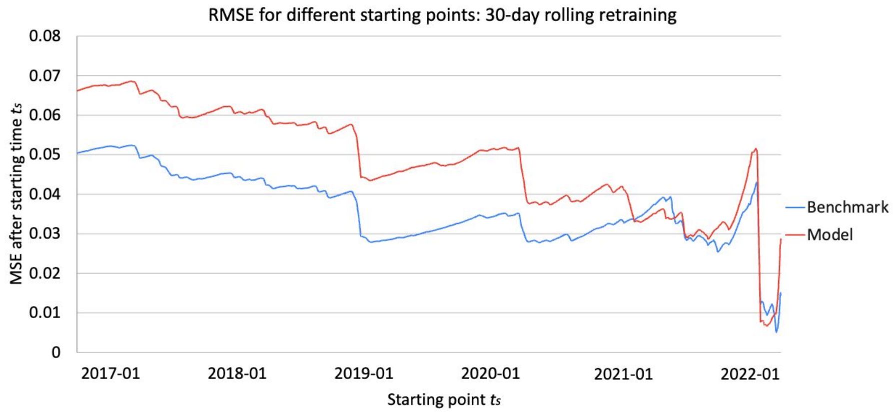
- Even though not better most of the time, can still use in combination with check on benchmark – possibly better together

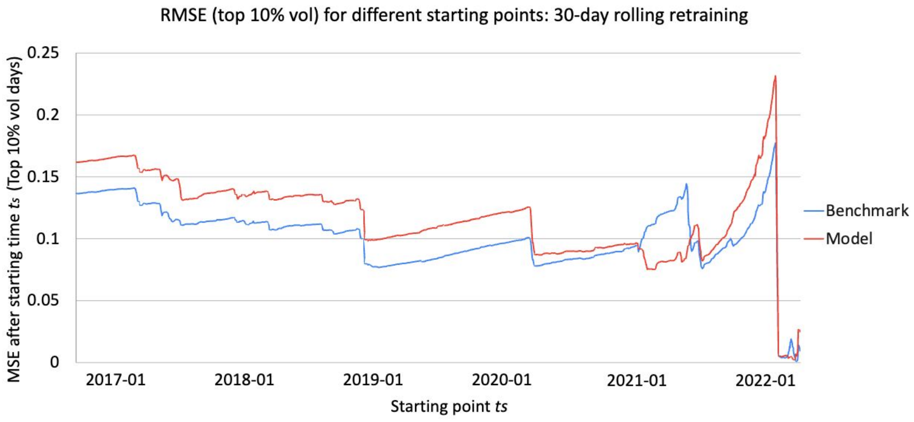
Model retraining periods:	1-day	7-day	30-day
Model RMSE	7.82%	18.83%	18.98%
Benchmark RMSE	3.77%	9.39%	19.80%
Model (top 10% vol) RMSE	15.41%	23.15%	29.84%
Benchmark (top 10% vol) RMSE	7.5%	12.13%	36.61%

Table 2: RMSEs of the models compared to benchmarks over the last year of the dataset (May 2021 - May 2022).

- Focus on last year of dataset (most training data)
- 30-day model can be better than benchmark, but sensitive to this choice


Difference in Squared Error: Benchmark - Model for 30-day rolling retrainings
0.40
0.30
0.20
0.10
0.00
-0.10
-0.20
-0.30
-0.40
-0.50
-0.60
-0.70
DSE
2017-01
2018-01
2019-01
2020-01
2021-01
2022-01


RMSE for different starting points: 30-day rolling retraining
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
MSE after starting time ts
2017-01
2018-01
2019-01
2020-01
2021-01
2022-01
Starting point ts
Benchmark
Model


RMSE (top 10% vol) for different starting points: 30-day rolling retraining
0.25
0.2
0.15
0.1
0.05
0
MSE after starting time ts (Top 10% vol days)
2017-01
2018-01
2019-01
2020-01
2021-01
2022-01
Starting point ts
Benchmark
Model

Conclusion

- Noisy signal, some price information recoverable
- At current state, not very actionable in practice (low quality, high complexity)
- Circuit breakers on oracle changes may get most of the gain

- Important area: DeFi depends on oracle prices but they’re often taken at face value.

Further research topics

- Incorporating further data sources
  - Difficulties: modeling approach, accessing some chain data within EVM, factoring manipulability into model
- Modeling how this architecture affects incentives of oracle providers
  - Can model as capital structure models. Interchange oracle provider with governors in existing models: https://arxiv.org/abs/2006.12388 https://arxiv.org/abs/2109.08939
  - *Idea:* CPF adds constraints to these models that tend to increase incentive compatibility of oracle provider

Economically securing oracle networks

- Current stake-slashing criteria for oracle networks
  - Suffers beauty pageant problem (same as oracle problem)
  - Rely on consensus of other node operators, but the consensus is not provably correct

- Designing alternative criteria: an ‘optimistic’ version of CPF
  - Oracle node operators report both prices and {0,1} whether CPF conditions would be violated (without executing)
  - Anyone could prove if the latter was reported incorrectly ⇒ slash node operator

Conclusion

- Noisy signal, some price information recoverable
- At current state, not very actionable in practice (low quality, high complexity)
- Circuit breakers on oracle changes may get most of the gain

- Important area: DeFi depends on oracle prices but they’re often taken at face value.

Twitter: @aklamun