Shillometer

Problem Statement

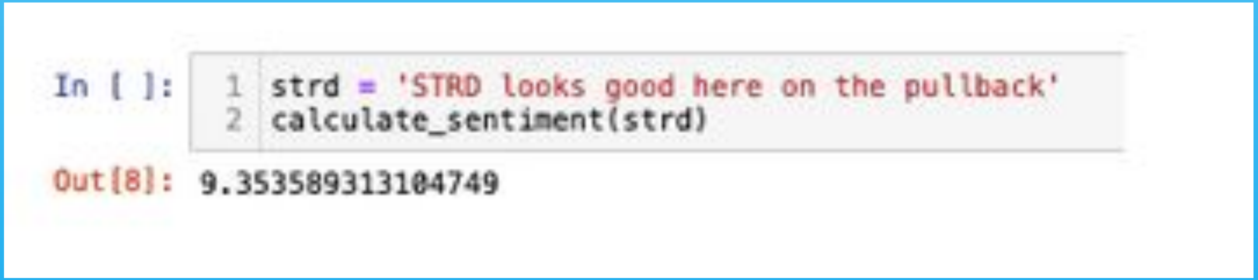
Crypto markets are heavily driven by sentiment and narratives. Explore scraping social media to get a sense for when these narratives might be building, in order to identify trade opportunities.

To do this, scrape high volume public Telegram groups and perform sentiment analysis on the results, then use that data as a feature within a number of machine learning m models.

If any of these models show promise, build a trade strategy and backtest this over the last 3-4 years, when. most of the major focus coins in today's market went live or saw their largest periods of price appreciation.

Data Collection

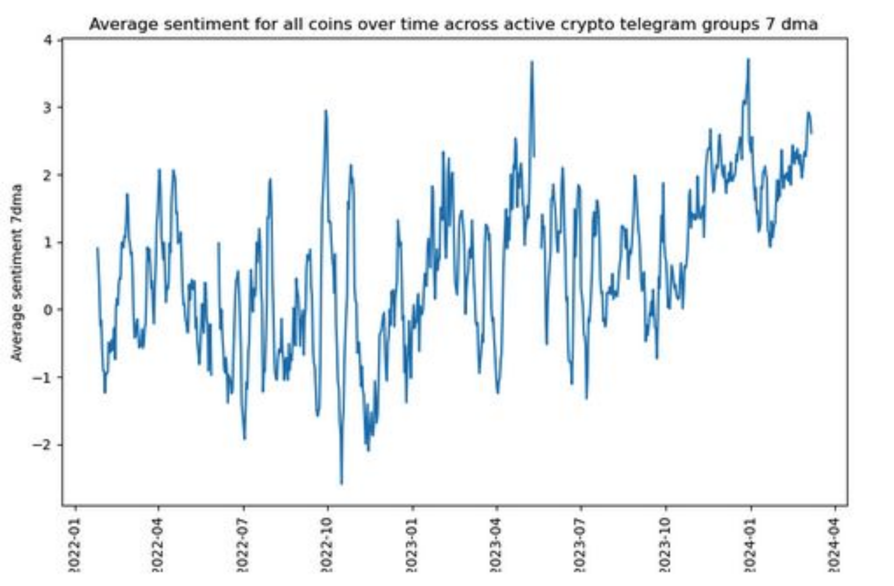
- I scraped 12 high volume public telegram chats for ~200k comments over the last 2 years, of which ~30k related to a specific coin
- I then ran this through a Hugging Face library for sentiment scoring from -10/+10

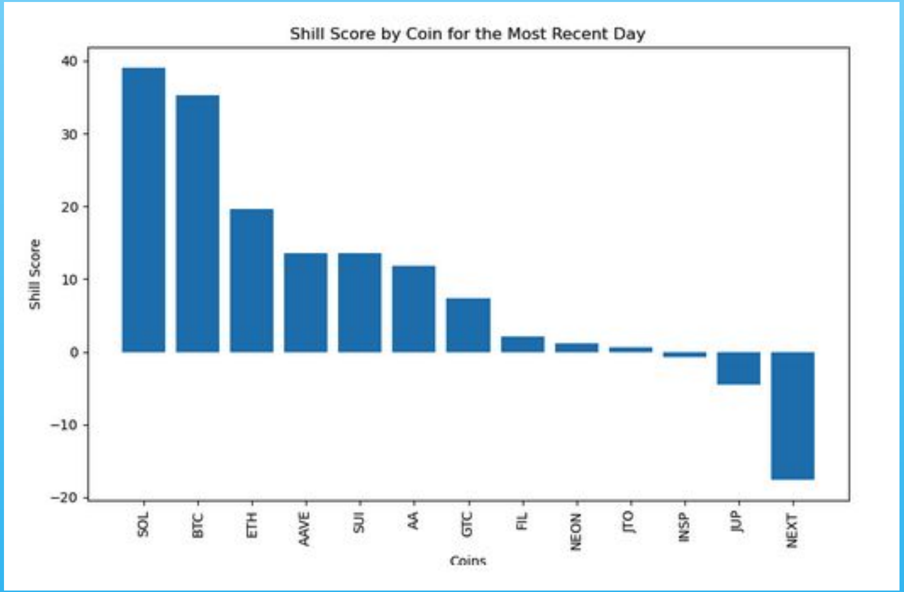


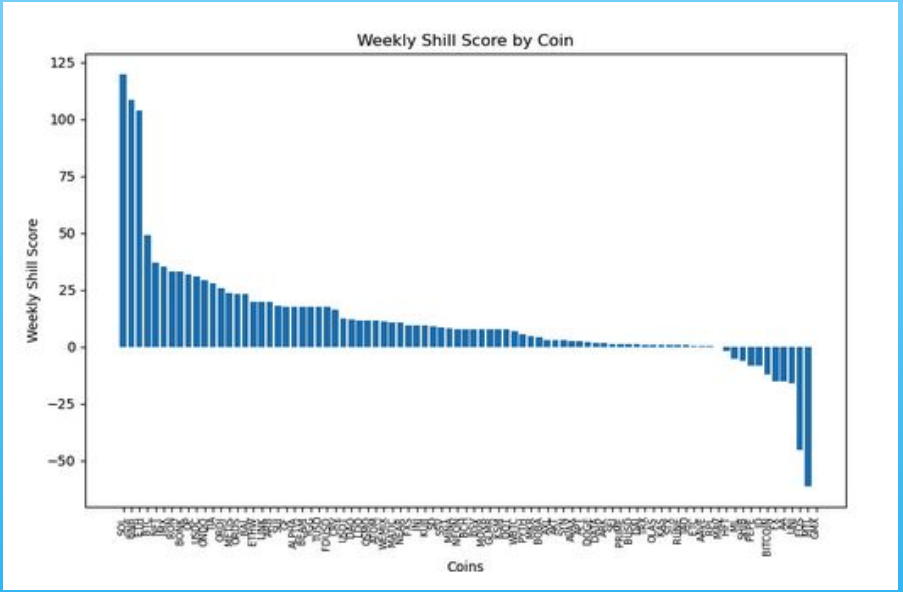
```
In [ ]:  strd = 'STRD looks good here on the pullback'
         calculate_sentiment(strd)

Out[8]:  9.353589313104749
```

EDA

- I examined the accuracy manually
- The comments were broadly correct through many were clustered around -3 to +3. In general though, the higher conviction comments were broadly accurate.

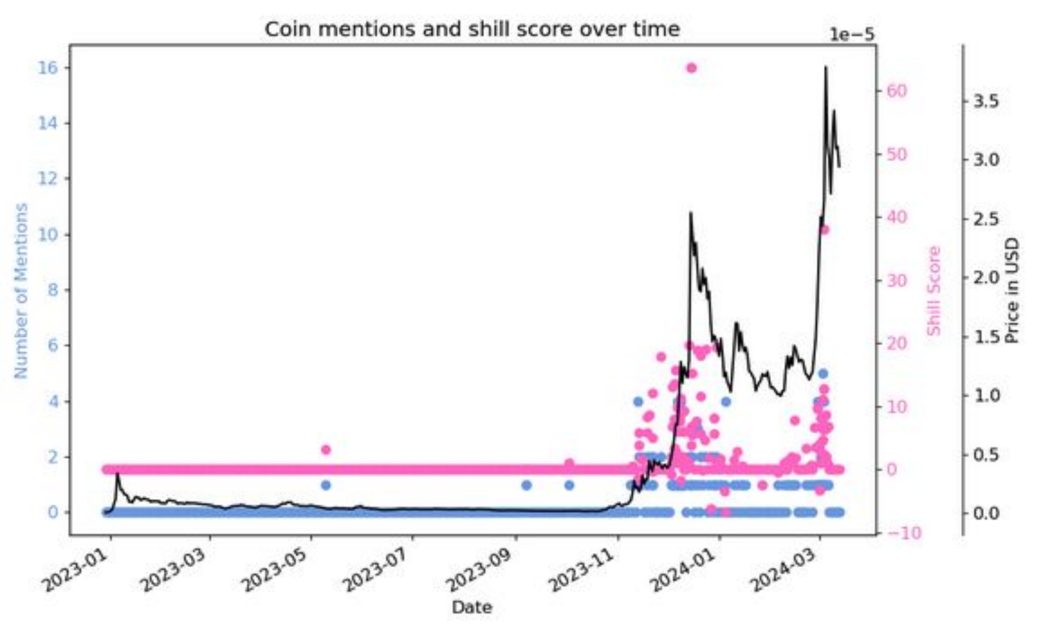



Shill Score by Coin for the Most Recent Day
Shill Score
40
30
20
10
0
−10
−20
SOL
BTC
ETH
AAVE
SUI
AA
GTC
FIL
NEON
JTO
INSP
JUP
NEXT
Coins


Weekly Shill Score by Coin
Weekly Shill Score
125
100
75
50
25
0
−25
−50
SOL
BITCOIN
Coins

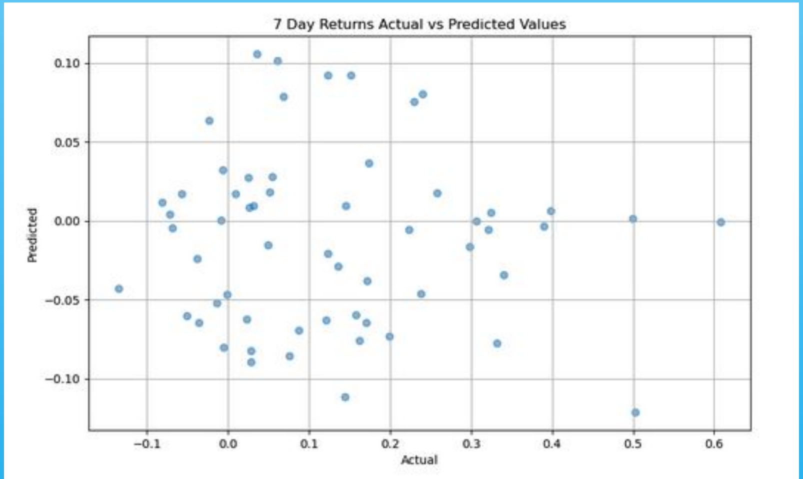
EDA

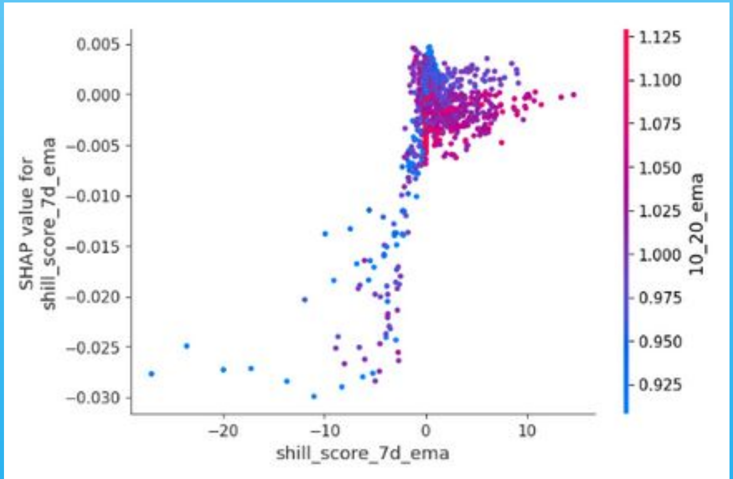
- I then combined this sentiment data with price history to see how ‘shill score’ ie [sentiment*number of comments per day], essentially my proxy for hype, correlated to price and whether there might be any leading indicator



Modelling

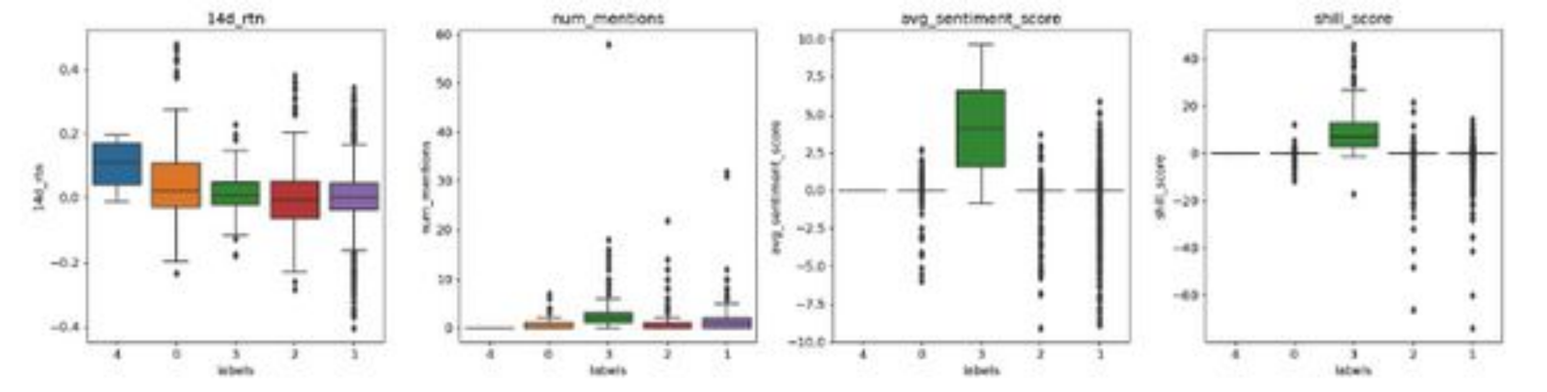
- I then tries to build a regression model, using RF and SVR. The results were not great, though SHAP analysis was somewhat more encouraging





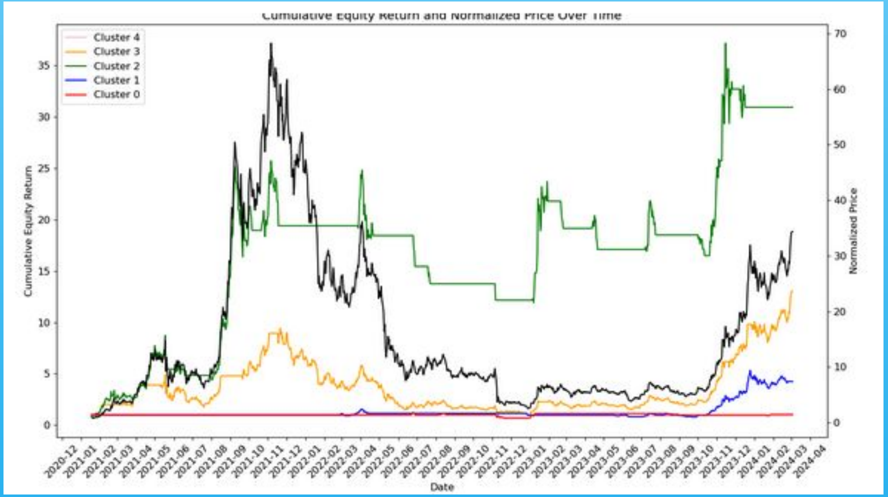
K-means and PCA

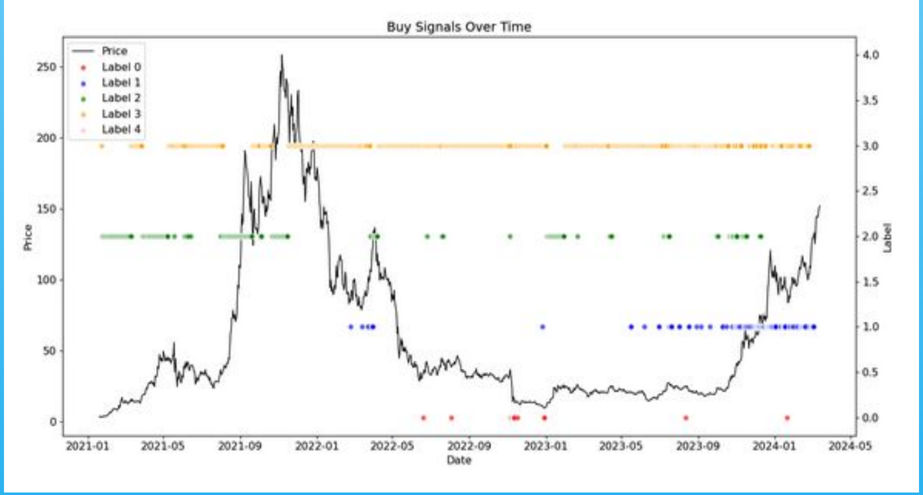
- Which brought me to clustering - on the basis that the stars might align across a number of technical factors at local points, while performing poorly in linear regression style approaches
- I used PCA for dimensionality reduction, in an effort to find tighter clusters and therefore cleaner labelling
- PCA didn’t really improve silhouette scores but I did seem to get better results on backtesting



Building a strategy

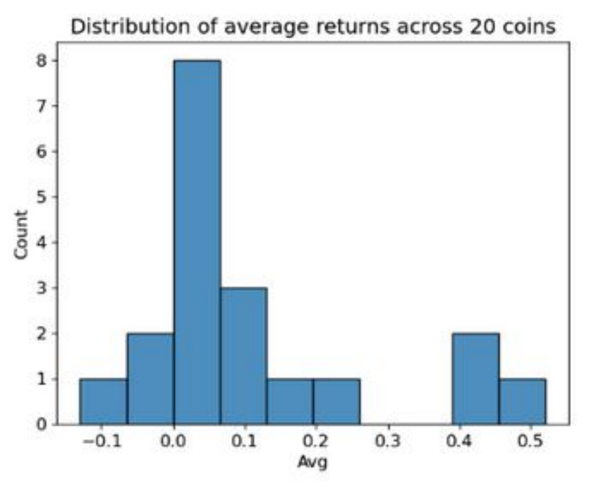
- I built a strategy of buying whenever the cluster trigger fired each day and holding for 14 days or cutting at a 10% stop loss





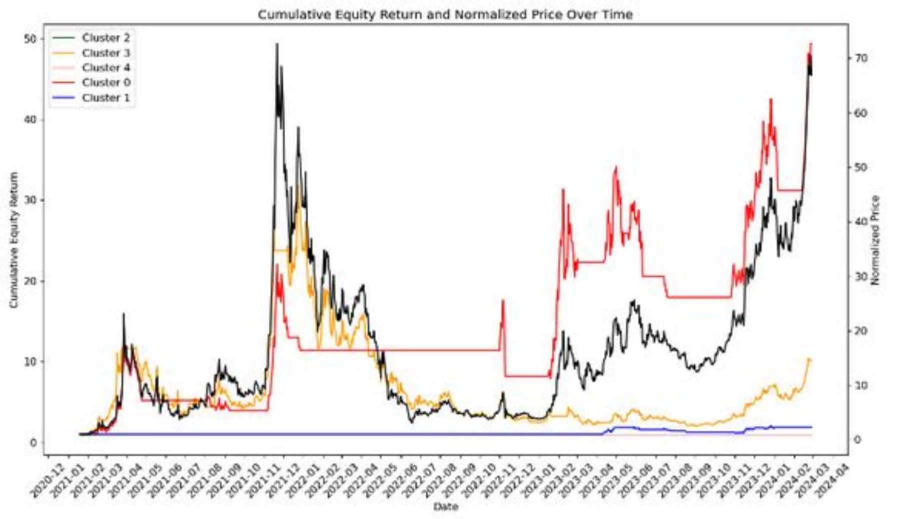
Backtesting across 20 coins

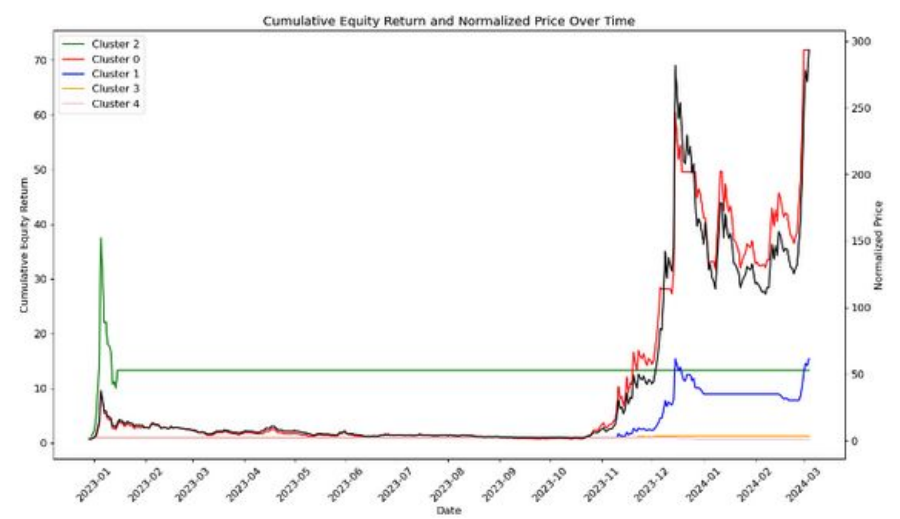
- I backtested the strategy more rigorously across 20 major coins and the results were fairly impressive - 160% return over 2 years and an average trade return of 10%



Conclusion

- The social media data was interesting in identifying hot, well hyped coins, but seemed to have limited predictive ability
- However, the wider clustering approach generally identified early momentum clusters, and was invested during aggressive market runs
- I think there is value in this approach and it could be used to help investors identify early buying opportunities


Cumulative Equity Return and Normalized Price Over Time
Cluster 2
Cluster 3
Cluster 4
Cluster 0
Cluster 1
Cumulative Equity Return
Normalized Price
Date


Cumulative Equity Return and Normalized Price Over Time
Cluster 2
Cluster 0
Cluster 1
Cluster 3
Cluster 4
Cumulative Equity Return
Normalized Price
Date