NYU

Short-Term Cryptocurrency Price Prediction

Based on Aggregate Market Signals and Sentiments

PRESENTED BY Zijin Hu, Chenye Yuan, Hongyi Zheng

08/17/2022

Agenda

1. Motivation & Business Understanding
2. Data Preparation & Feature Selection
3. Modeling
4. Result/Evaluation
5. Future Works

NYU

Motivation & Business Understanding

01

NYU

Motivation

Why Crypto?

- High volatility – more signals, thus more trading opportunities
- Traded over multiple exchanges – provides more features
- Emerging asset class - more originality for our project

Business Understanding

- By **Efficient Market Hypothesis**, all publicly available information will be reflected on the price of an asset.
- Short-term trading is more of a **game between market agents**. The price of an asset in short term is more depending on the **market sentiments**, instead of fundamental or macro information. Thus, the signal offered by the market, could in return influence the market, which **blurs the boundary of leading and lagging indicators**.
- **Cryptocurrency, as an assets which value is determined by consensus**, short term fluctuation may be lead by the momentum that could be uncovered by mining market indicators. Meanwhile, the **high volatility** of cryptocurrency provide more signals and trading opportunity. The **exchange liquidity differences** also provide more features.
- Therefore, we choose some indicators that traditionally or logically correlated with price movement, and trying to predict the price directional movement of cryptocurrency in the short term (~5Min).

Data Preparation

02

NYU

Features

- Open Interest
- Top Trader Long/Short Ratio (Account)
- Top Trader Long/Short Ratio (Positions)
- Global Long/Short Ratio (Account)
- Taker Buy/Sell Volume
- Past Price Change

Price data is obtained from Yahoo Finance API, other data is obtained from Binance API

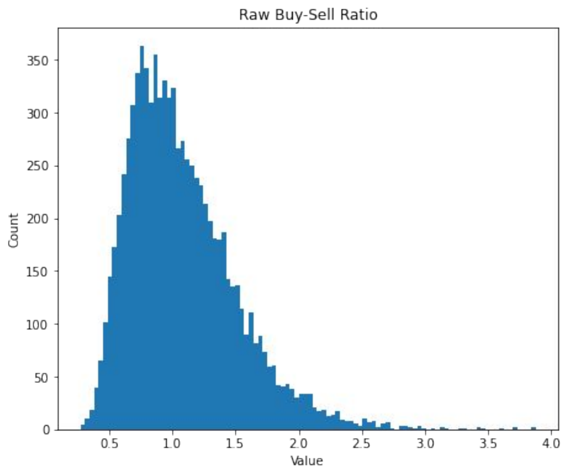
Response:

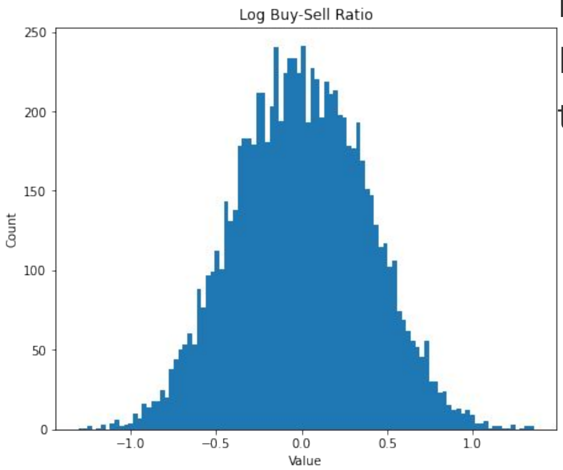

```
[
    {
            "buySellRatio":"1.5586",
            "buyVol": "387.3300",
            "sellVol":"248.5030",
            "timestamp":"1585614900000"

    },
```


	sumOpenInterestChange	logBuySellRatio	logTopTraderAccountsLongShortRatio	logTopTraderPositionsLongShortRatio	logGlobalAccountsLongShortRatio	pastPxChange	futurePxChange
2022-07-18 04:00:00	0.000595	-0.822345	-0.085231	0.152378	-0.151521	-0.000876	-0.000178
2022-07-18 04:05:00	0.000212	0.127601	-0.084034	0.151691	-0.154667	-0.000178	0.000494
2022-07-18 04:10:00	0.002121	-0.230420	-0.081644	0.151089	-0.156303	0.000494	-0.000905
2022-07-18 04:15:00	0.000715	-0.362693	-0.078394	0.151003	-0.153501	-0.000905	-0.002697
2022-07-18 04:20:00	0.001064	-0.367736	-0.079260	0.150487	-0.155485	-0.002697	0.002869
...	...	...	...	...	...	...	...
2022-08-16 23:30:00	-0.000127	-0.027474	0.598452	0.128833	0.610689	0.000051	-0.000469
2022-08-16 23:35:00	-0.000070	-0.003005	0.599770	0.129799	0.610689	-0.000469	0.000362
2022-08-16 23:40:00	-0.000218	-0.012478	0.596691	0.129448	0.608950	0.000362	0.000169
2022-08-16 23:45:00	-0.000025	0.469066	0.596691	0.128921	0.608079	0.000169	0.000758
2022-08-16 23:50:00	0.001618	0.017250	0.596251	0.129272	0.607644	0.000758	-0.000495

8587 rows × 7 columns

Feature Engineering

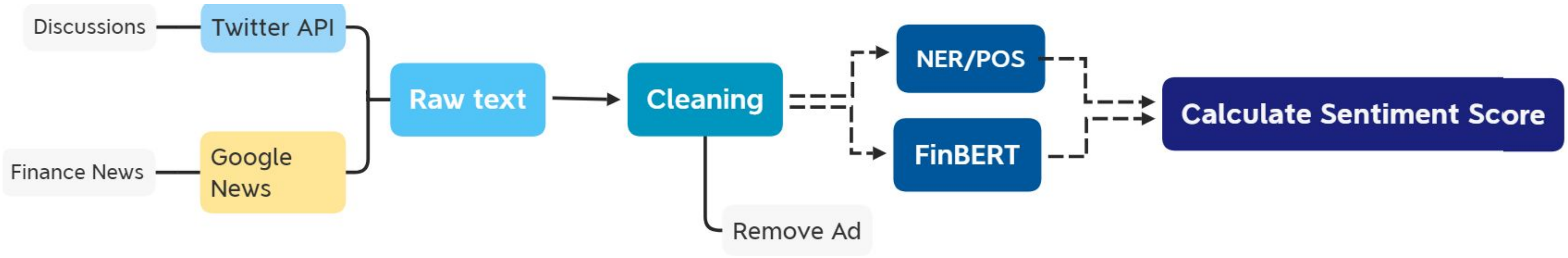




- Convert raw price / open interest data to their change in percentage space over time.
- Take log transformation to all ratio data to make our model more robust to outliers.
- Convert unix timestamp to datetime
- Normalize each feature to facilitate further operations. (subtract feature by mean and divide by standard deviation) Make every feature zero-centered so that we could regress without intercept.

Exploratory Data Analysis

Sentiment



NYU

Modeling

03

NYU

Linear Regression

- Simple yet powerful model
- Reduce the risk of overfitting
- The contribution of each feature is interpretable
- Compatible with some efficient feature selection algorithm.

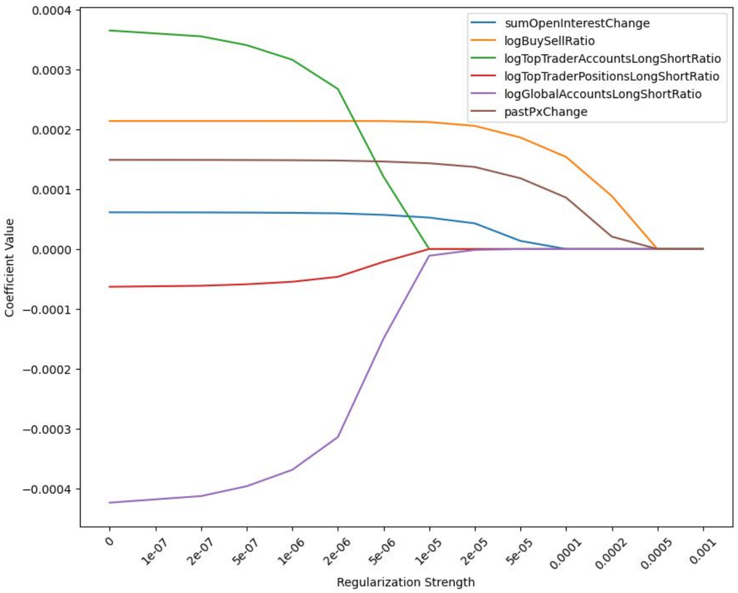
OLS Regression Results

Dep. Variable:	futurePxChange	R-squared (uncentered):	0.055
Model:	OLS	Adj. R-squared (uncentered):	0.054
Method:	Least Squares	F-statistic:	41.77
Date:	Tue, 16 Aug 2022	Prob (F-statistic):	9.03e-50
Time:	19:02:29	Log-Likelihood:	23265.
No. Observations:	4320	AIC:	-4.652e+04
Df Residuals:	4314	BIC:	-4.648e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
sumOpenInterestChange	0.0065	0.010	0.672	0.502	-0.012	0.025
logBuySellRatio	0.0003	5.09e-05	6.013	0.000	0.000	0.000
logTopTraderAccountsLongShortRatio	0.0002	0.000	0.578	0.563	-0.001	0.001
logTopTraderPositionsLongShortRatio	-0.0003	0.000	-1.106	0.269	-0.001	0.000
logGlobalAccountsLongShortRatio	-2.033e-05	0.000	-0.057	0.955	-0.001	0.001
pastPxChange	0.1593	0.017	9.283	0.000	0.126	0.193

NYU

Feature Selection — Lasso



- For linear regression, Lasso could be used to conduct feature selection. It will squeeze the coefficients of unhelpful/highly correlated features to zero.
- After applying regularization in Lasso regression, we select features with non-zero coefficients and use them to refit the model.
- For each coin, grid search the best regularization strength.

Evaluation

04

Pipeline

			insample	outOfSample	PnL
coin	regularization strength	date			
BTC	0.000	2022-08-03	0.041329	0.034273	587.212310
		2022-08-04	0.042241	0.073870	907.415144
		2022-08-05	0.043744	0.073724	1113.635126
		2022-08-06	0.046352	0.025732	142.719724
		2022-08-07	0.045462	-0.065738	342.969880
...	...	...	...	...	...
DOT	0.001	2022-08-12	0.000000	0.000000	0.000000
		2022-08-13	0.000000	0.000000	0.000000
		2022-08-14	0.000000	0.000000	0.000000
		2022-08-15	0.000000	0.000000	0.000000
		2022-08-16	0.000000	0.000000	0.000000

1232 rows × 3 columns

```python
if __name__ == '__main__':

    coins = ["BTC", "ETH", "ADA", "SOL", "LTC", "DOGE", "XRP", "DOT"]

    with Pool() as pool:
        results = pool.imap_unordered(get_data_for_coin, coins)
        result_df = pd.concat(results, keys=coins)
        result_df.to_csv("result_by_coin.csv")
```

- We wrote a python script that makes use of the power of multiprocessing to search the out-of-sample r2 performance for every coin-hyperparameter combination.
- Use rolling validation to get r2 data by date: for each **d**, use data from day **d-15** to **d-1** to fit the model and use that model to make prediction on day **d**
- Concatenate the results into one large dataframe and conduct the result analysis.

NYU

Results

- Find the regularization strength that achieves highest out-of-sample r2
- Convert predictions to actual trading performance.
- Compute sharpe value (PnL mean / PnL std)

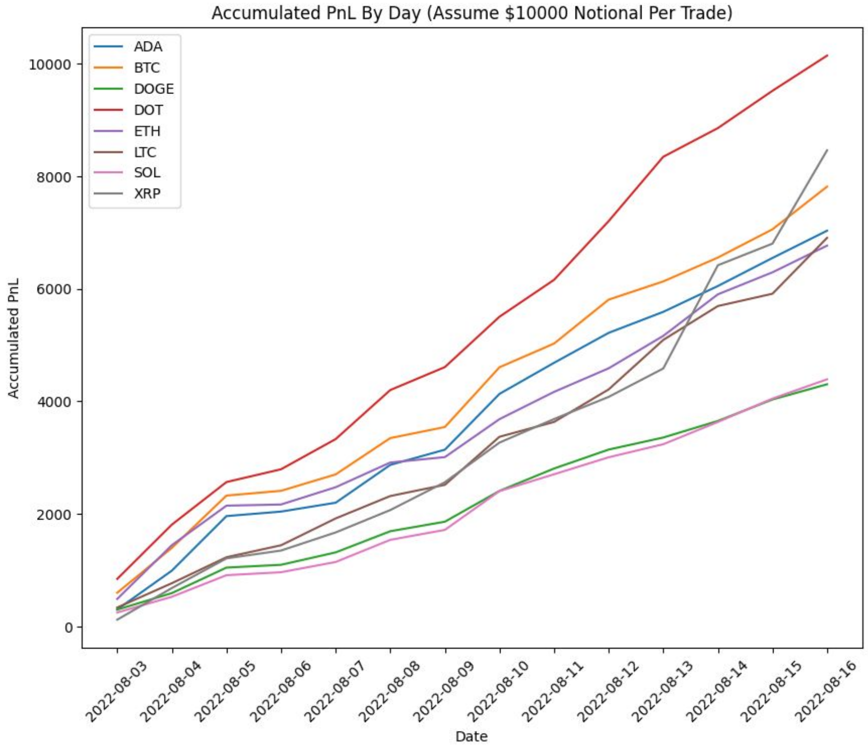
```
coin
ADA     0.082766
BTC     0.048720
DOGE    0.057387
DOT     0.095938
ETH     0.086001
LTC     0.068635
SOL     0.053608
XRP     0.053043
Name: outOfSample, dtype: float64
```

```
coin
ADA       (ADA, 0.0001)
BTC       (BTC, 0.0001)
DOGE     (DOGE, 0.0001)
DOT       (DOT, 0.0001)
ETH       (ETH, 0.0001)
LTC        (LTC, 2e-05)
SOL        (SOL, 5e-05)
XRP       (XRP, 0.0001)
Name: outOfSample, dtype: object
```

```
coin  regularization strength
ADA   0.00010                    1.876568
BTC   0.00010                    1.947536
DOGE  0.00010                    2.469474
DOT   0.00010                    2.836343
ETH   0.00010                    1.961242
LTC   0.00002                    1.897977
SOL   0.00005                    2.114934
XRP   0.00010                    1.189168
Name: sharpe, dtype: float64
```

Results

Rolling Validation Result By Day
Out of Sample R^2
0.15
0.10
0.05
0.00
−0.05
ADA
BTC
DOGE
DOT
ETH
LTC
SOL
XRP
2022-08-03
2022-08-04
2022-08-05
2022-08-06
2022-08-07
2022-08-08
2022-08-09
2022-08-10
2022-08-11
2022-08-12
2022-08-13
2022-08-14
2022-08-15
2022-08-16
Date



Accumulated PnL By Day (Assume $10000 Notional Per Trade)
Accumulated PnL
10000
8000
6000
4000
2000
0
ADA
BTC
DOGE
DOT
ETH
LTC
SOL
XRP
2022-08-03
2022-08-04
2022-08-05
2022-08-06
2022-08-07
2022-08-08
2022-08-09
2022-08-10
2022-08-11
2022-08-12
2022-08-13
2022-08-14
2022-08-15
2022-08-16
Date

Future Work

05

NYU

Limitations

"Bad" Data Source:

- Unlike news headlines, tweets or reddit posts' sentiment can be ambiguous or objectively neutral, leading to poor accuracy.
- Difficulties in extracting valuable opinions and information from other kind of texts.
- On smaller time scales, whether reddit and twitter posts are sufficiently correlated with price movements remains to be further verified

Possible Solution:

- Instead of collecting all tweets, try ranking tweets based on viewer's rating (like/dislike) and authority (number of followers)
- Explore more time sensitive social media data source (e.g. Telegram)

Potential Future Work

- There is no Holy Grail. More alternative data sources could be used.
- Fine-tune the power adjustment to some of the features
- Adding more models to our pipeline.
- Take commission fee into consideration during evaluation stage.
- Current model is applied toward all data and time without selection. For deployment, besides directional prediction, we need better model for timing selection to reduce noise and increase accuracy. SL/TP strategy is also required. A potential solution is to apply decision tree to select entering time and position.

END