Problem
To check whether volume spike is a good predictor of daily closing price

Preparations

1. Raw data
    a. 2 years data of IDX30 stocks from 2022-07-01 09:00:00+07:00 to 2024-04-05 09:00:00+07:00
    b. Data granularity is hourly
    c. Sample raw data on ACE HARDWARE

timestamp	open_aces	close_aces	volume_aces
2024-06-28 12:00:00+07:00	860	855	826700
2024-06-28 13:00:00+07:00	860	855	826700
2024-06-28 14:00:00+07:00	860	855	826700
2024-06-28 15:00:00+07:00	855	855	3038000
2024-06-28 16:00:00+07:00	855	855	0

2. Data Processing
    a. Calculated daily percent change and hours to market close
    b. Split data to train and test for further testing
    c. Define target rules, eg if daily percent change is greater than 0.1%
    d. Normalize every feature using interquartile scaler (robust scaler) except temporal data
    e. Calculate rolling standard deviation of every predictors. In this case, the calculation is rolling every 25 days * 7 hours
    f. Define the volume spike, eg if spike is greater than the 1 * rolling volume standard deviation
    g.

Check Statistical Significance
We use Pearson's correlation to quantify the link between volume traded and *daily* (not hourly) close prices. Student's T-Test to check whether the spike population is sufficiently different with the non-spike population

The hypothesis is as follows
Null Hypothesis (H0): There is no significant difference in the daily percentage change between instances with and without volume spikes.
Alternative Hypothesis (H1): There is a significant difference in the daily percentage change between instances with and without volume spikes.

Below is the result, sorted by T-test significance

Asset	T-test P-Values	Correlation Values	Volume Mean	Volume Median	Volume 95 Percentile	Number of Spikes
arto	0.000000	0.282752	3,328,655	1,973,400	11,463,000	611
brpt	0.000000	0.289413	14,917,927	7,866,900	44,647,135	684
inkp	0.000000	0.201046	986,603	667,150	3,061,310	578
medc	0.000000	0.196492	14,529,043	8,691,300	46,114,600	623
antm	0.000006	0.181769	9,472,761	5,729,350	30,786,460	566
icbp	0.000793	0.098729	885,463	583,900	2,625,850	640
cpin	0.002465	-0.013869	1,075,527	680,500	3,313,050	639
akra	0.003100	0.092754	5,332,371	3,346,550	17,770,520	695
aces	0.004279	0.295838	11,480,693	6,236,350	39,005,505	628
smgr	0.004831	0.169362	1,368,529	947,700	3,958,745	614
pgeo	0.017746	0.231160	8,331,198	2,716,050	33,836,785	476
klbf	0.029187	0.038454	4,850,125	3,298,000	14,459,060	552
buka	0.055882	0.257300	31,045,901	16,345,900	104,605,085	603
itmg	0.060912	-0.067552	380,724	234,350	1,185,750	548
mdka	0.079074	0.021241	7,153,547	4,868,600	19,935,290	656
untr	0.096887	-0.040517	759,889	572,650	1,988,165	583
bbca	0.130758	0.198885	9,780,250	7,627,300	24,299,400	635
adro	0.191627	0.030467	9,750,783	6,570,400	28,891,935	576
bbri	0.250236	0.061770	17,781,225	13,680,850	46,315,820	671
indf	0.325400	0.006314	1,108,393	839,900	2,938,400	640
bbni	0.380460	0.066083	4,859,875	3,573,650	13,440,195	696
unvr	0.394552	-0.065028	2,210,846	1,461,000	6,248,850	563
pgas	0.547042	0.117721	9,188,266	5,394,200	28,757,110	560
goto	0.718049	0.143709	354,748,174	187,603,200	1,226,965,600	691
asii	0.784636	0.049036	6,889,974	5,175,300	18,590,130	612
ptba	0.794795	-0.056529	3,609,880	1,975,150	10,394,550	525
tlkm	0.805706	-0.074869	13,282,817	9,977,000	36,074,240	610
inco	0.870530	0.105829	1,823,668	1,081,450	6,207,265	608
bmri	0.955103	0.040701	9,368,169	6,866,450	25,487,505	722
amrt	0.966565	-0.040311	4,200,943	2,495,450	12,953,215	598

Below is the result, sorted by Pearson's correlation

Asset	T-test Values	Correlation Values	Volume Mean	Volume Median	Volume 95 Percentile	Number of Spikes
aces	0.004279	0.295838	11,480,693	6,236,350	39,005,505	628
brpt	0.000000	0.289413	14,917,927	7,866,900	44,647,135	684
arto	0.000000	0.282752	3,328,655	1,973,400	11,463,000	611
buka	0.055882	0.257300	31,045,901	16,345,900	104,605,085	603
pgeo	0.017746	0.231160	8,331,198	2,716,050	33,836,785	476
inkp	0.000000	0.201046	986,603	667,150	3,061,310	578
bbca	0.130758	0.198885	9,780,250	7,627,300	24,299,400	635
medc	0.000000	0.196492	14,529,043	8,691,300	46,114,600	623
antm	0.000006	0.181769	9,472,761	5,729,350	30,786,460	566
smgr	0.004831	0.169362	1,368,529	947,700	3,958,745	614
goto	0.718049	0.143709	354,748,174	187,603,200	1,226,965,600	691
pgas	0.547042	0.117721	9,188,266	5,394,200	28,757,110	560
inco	0.870530	0.105829	1,823,668	1,081,450	6,207,265	608
icbp	0.000793	0.098729	885,463	583,900	2,625,850	640
akra	0.003100	0.092754	5,332,371	3,346,550	17,770,520	695
bbni	0.380460	0.066083	4,859,875	3,573,650	13,440,195	696
bbri	0.250236	0.061770	17,781,225	13,680,850	46,315,820	671
asii	0.784636	0.049036	6,889,974	5,175,300	18,590,130	612
bmri	0.955103	0.040701	9,368,169	6,866,450	25,487,505	722
klbf	0.029187	0.038454	4,850,125	3,298,000	14,459,060	552
adro	0.191627	0.030467	9,750,783	6,570,400	28,891,935	576
mdka	0.079074	0.021241	7,153,547	4,868,600	19,935,290	656
indf	0.325400	0.006314	1,108,393	839,900	2,938,400	640
cpin	0.002465	-0.013869	1,075,527	680,500	3,313,050	639
amrt	0.966565	-0.040311	4,200,943	2,495,450	12,953,215	598
untr	0.096887	-0.040517	759,889	572,650	1,988,165	583
ptba	0.794795	-0.056529	3,609,880	1,975,150	10,394,550	525
unvr	0.394552	-0.065028	2,210,846	1,461,000	6,248,850	563
itmg	0.060912	-0.067552	380,724	234,350	1,185,750	548
tlkm	0.805706	-0.074869	13,282,817	9,977,000	36,074,240	610

It appears that volume spike is a weak predictor for daily closing prices, as even the strongest correlation, exemplified by ACES, only reaches a correlation coefficient of 0.29. Moreover, the significance between populations with spikes versus no spikes varies greatly across different assets. For instance, in green assets, the close price population of those with spikes and non spikes are statistically different, whereas in red assets, the distribution appears entirely random.

Notably, most of the green assets, for example CPIN, exhibits a very good t-test metric but a low absolute correlation coefficient. This discrepancy may imply that while there is a statistical distinction between closing prices vs volume spikes, it may not translate towards a decisive movement in price

Modeling

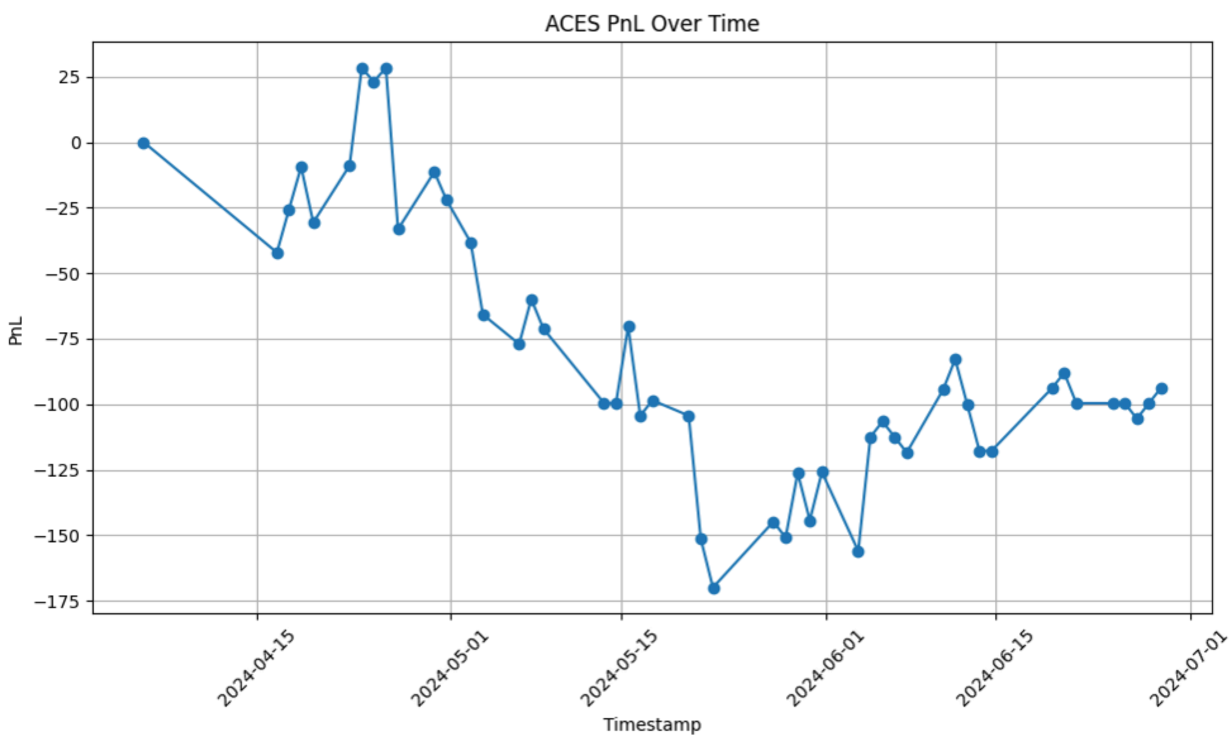
Nevertheless, we attempt to employ an LSTM deep learning model to predict closing prices. We input sequences consisting of 25 days * 7 hours of previous closing data and the latest volume as predictors, aiming to demonstrate the relationship between volume and price. The model utilizes these predictors to forecast whether the daily percent price change exceeds 0.1%.

Initially, our focus is on the extreme cases of ACES, which exhibits the highest volume-price correlation.

	Predicted Negative ACES	Predicted Positive ACES
Actual Negative ACES	72.56	3.13
Actual Positive ACES	17.12	7.19

For ACES, the model demonstrates reasonable ability to predict percent changes. Out of every 100 predictions, it correctly forecasts that the price will not exceed the 0.1% limit approximately 72 times. It also accurately predicts instances where the price exceeds this threshold about 7 times. However, in approximately 17 instances, the model predicts the price will stay below the threshold, but it actually exceeds it, resulting in missed trading opportunities. Conversely, around 3 times, the model predicts the price will rise when it does not, potentially leading to trading losses.

Now we try to bring the model to life using dataset that it has not seen yet, opening up position when it predicted up, and close by the end of day

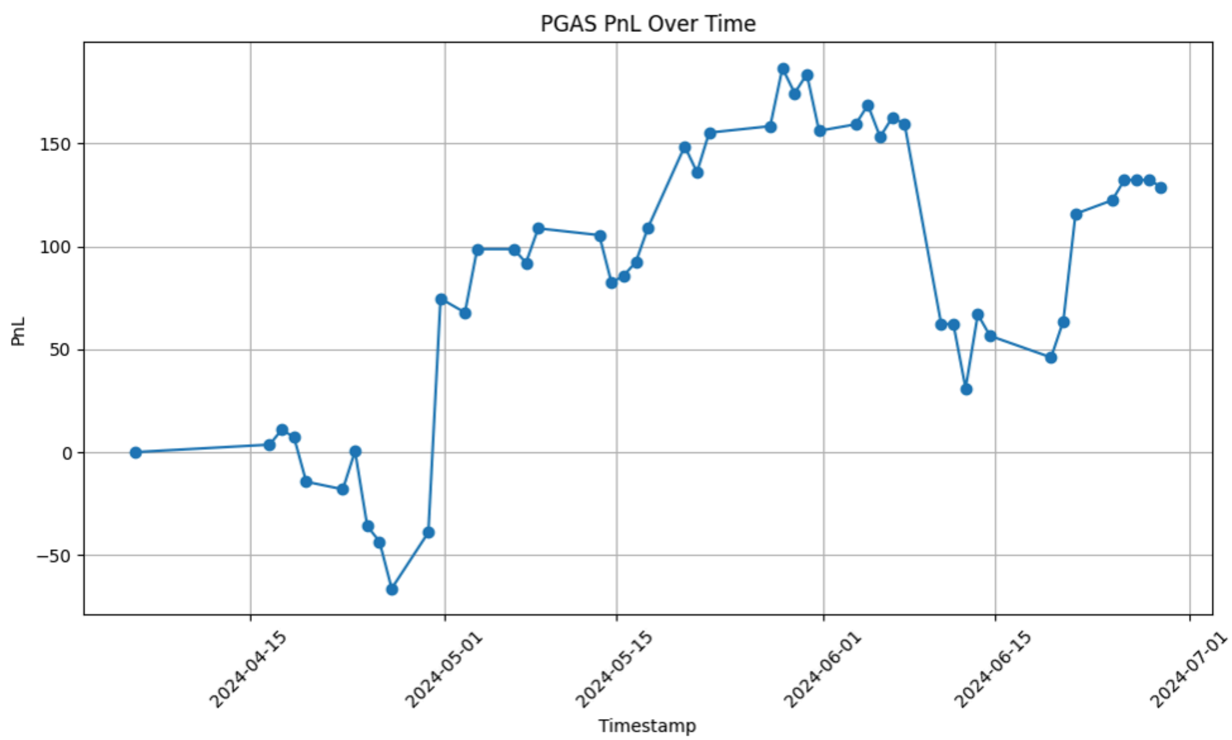


Over the 3-month period, a majority of trades have yielded positive results. This, coupled with a favorable confusion matrix, suggests that using price and volume as predictors could be a promising approach. However, it's notable that when the model predictions are incorrect, the resulting drawdowns are quite substantial. Therefore, there is a clear need to minimize errors in the model.

On the other hand, the best performing asset is PGAS

	Predicted Negative PGAS	Predicted Positive PGAS
Actual Negative PGAS	82.81	0.63
Actual Positive PGAS	11.67	4.90

For PGAS, the model demonstrates reasonable ability to predict percent changes. Out of every 100 predictions, it correctly forecasts that the price will not exceed the 0.1% limit approximately 83 times. It also accurately predicts instances where the price exceeds this threshold about 6 times. However, in approximately 11 instances, the model predicts the price will stay below the threshold, but it actually exceeds it, resulting in missed trading opportunities. Conversely, around 0.63 times, the model predicts the price will rise when it does not, potentially leading to trading losses.



While the model performs well on PGAS, it shows a notably low T-test score of 0.547042 and a modest correlation of 0.117721. This raises concerns that the LSTM model may primarily emphasize the predictive power of previous closing prices rather than volumes, challenging our assumption of a volume-price correlation.

Conclusion

In summary, our analysis suggests that there is insufficient evidence to reject the hypothesis that there is no relationship between volume and volume spikes with most of the asset's closing prices, as indicated by both statistical tests and predictive modeling.

However, several notable flaws in our experiments highlight areas for improvement. The primary limitation is the lack of data, both in terms of asset selection and time frame. To address this, we could expand our dataset by including more historical data prior to 2022. It's important to note that extending the dataset to include the Covid-19 period may introduce biases due to unique market conditions at that time, which is why we initially avoided its inclusion.

Additionally, another significant flaw is the uniform application of parameters, statistics, and models across all assets. Moving forward, it's essential to analyze each asset independently, considering that their behaviors may vary significantly.