ETH Price Predicting by Historical Data and On-chain Information

Hengyu Zhou

May 2nd, 2022

Contents

1 Abstract

My project was inspired by the observation of co-movement in crypto trading market(ie. when BTC price surges/jumps, ETH price would follow the trend shortly). I aim to discover relations between ETH and BTC by investigate in price movement and on-chain information. By closely examining minute level and daily level data on block-chain network related to these assets, we may find out the price relation between different tokens and may simplify our investment strategy on crypto by connecting everything with BTC. This conversion enables us to deflect valuation risks compares to treat them as individual investment and control risk seperately.

By building up the relation, we may also build up trading strategy between theses trading pairs in the future. According to any change in the market price/on-chain activity we can forecast the price trend in underlying asset, thus giving informed traders an first-mover advantage to either speculate or hedge their exposure accordingly while the rest of the market participants are still absorbing and reacting to this event. Our research exploits the inefficiency in the markets related to delayed volatility response and and has the potential of making significant arbitrage profit.

2 Introduction

Cryptocurrencies have evolved to become one of the most trending topics in recent economic and financial issues. Since Dotcom crisis, internet sales are booming with more and more tech savvy consumers go online shopping. Online commerce was mainly intermediated by financial institutions serving as trusted third parties to process electronic payments. Despite this system was capable enough for most transactions, it has a lot of limitations due to controls imposed by financial institutions. Thus, there is a need in the market for a decentralized currency that not only bypass financial controller, but also is cheap and anonymous. Thus, cryptocurrencies are created as a subset of virtual currencies that use cryptography for security. These are decentralized and open source currencies that can serve on a peer-to-peer basis, which perfectly fits the need from the market. The creation of cryptocurrencies rely on solving very complex cryptographic algorithm which are conducted through connected network of computers performing computationally expensive mathematical operations. Currently there are more than 2175 different types of cryptocurrencies been developed and they feature certain computer protocols that are out of any government control. As a result, these currencies have aggressive swings in their prices as it is largely based on public perceptions rather than hinged on the values of certain tangible assets. It is therefore very hard to make related risk assessment. With the increase of the pries of cryptocurrencies, mining has also turned into a very profitable business for the people.

I modeled the correlation between coins and intend to price one coin using the on-Chain information regarding to another coin, Bitcoin. The relative val-

uation method deflects the risk in valuing a crytocurrency directly in absolute dollar term and captures the similarity sharing among coins in terms of computer protocols and mining procedures. Through constructing valuation chains from Bitcoin to ETH in the market, my model can precisely quantify the collateral impact made by an volatility event on Bitcoin to ETH in the markets dynamically. This information would be very advantageous to coin traders as they can use the model to discover the intrinsic value of coins of their interest under the event and make informed decisions.

3 Conceptual Framework

3.1 Basic idea

3.1.1 Price Relation

Predicting BTC price is hard. Researchers have been deploying different kind of complicated models to depict the secret of this new popular asset. But as more and more individual investors/ garage fund entering this space, the BTC price is harder and harder to predict as it's still not a rational market. But if you look around on all the other crypto assets like ETH, you may find out though it have its own reason to rise in price, still it somehow linked to BTC price movement. Everytime BTC jumps or surges, you would instantly notice that ETH would react in a short time(1 min or 2) and follow the market trend.

I discovered this kind of trend in September 2021 when I was actively trading in the market but never got the chance to testify and model this observation. But after taking this class, I think I finally got the chance and the tool to explore the problem.

One possible reason is that BTC takes over 40% of the total marketcap so even the slightest move would also trigger chain reaction in the market. But this also means we use BTC as a anchor to build up trading portfolios around it. We can use the time lag to predict the price of other asset when there is signal showing up from bitcoin, buy in before it rise, sell out before it jumps and in the long run we will establish a system calculating ROI on BTC. Which means as long as you believe in BTC spirit and its development, you would get abundant revenue along this road.

One simple example is, there is a pattern that when BTC changes largely in price, ETH investor would follow the trend to buy or sell, but it takes time. Which would leave us with a time window to trade our asset. If BTC rises, ETH would also rise in one minutes, then you can sell BTC and buy ETH — so when ETH rises and reaches the balance point, you have more money than you do before the trade. If BTC falls, ETH would also falls in one minutes, if you sell ETH and buy BTC right now, then you would got more BTC than you used to. As long as you are calculating ROI by how many BTC you get at the end, you would be happy with this kind of strategy.

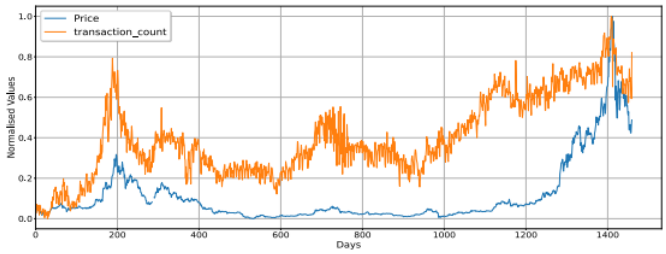


Figure 1: Correlation between normalized prices and transaction count

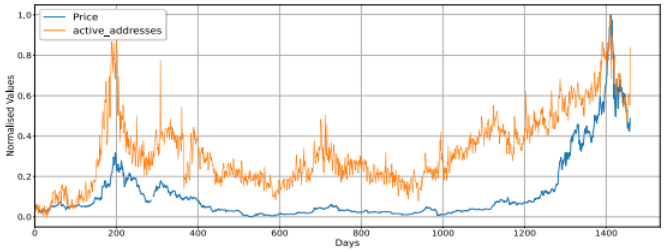
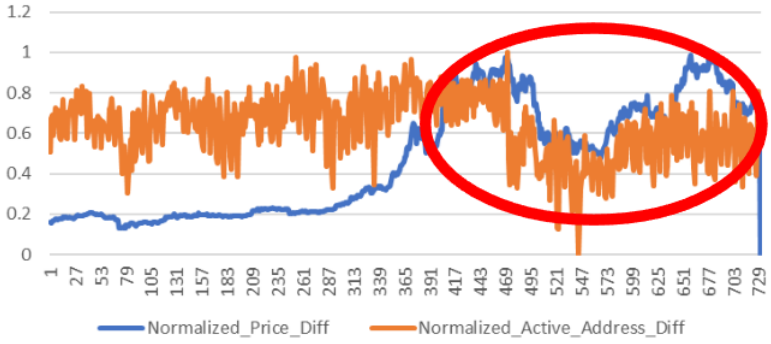


Figure 2: Correlation between normalized prices and active addresses



F

Figure 3: Correlation between normalized price differences and active addresses count differences

3.1.2 On-Chain Analysis Approach

As I captured the opportunity from the price relation, it doesn't mean we will always win. At the end of the day the price of an asset will definitely relate to itself. There must be some metrics of ETH influencing its appearance when BTC stays calm. In the past two years we have seen several coins make it to the top 10 and then left the list regardless of how BTC shows. There should be an explanation.

On-chain metrics is information generated by the blockchain network, such as the size of the blockchain, the number of blocks attached to it, or the difficulty of mining blocks. Their purpose is to inform interested parties about the state of a blockchain network; this data inherits the properties of these networks, such as their inherent transparency, tamper-resistant and decentralized nature.On-chain data is often recorded in a time-series manner, each metric offers insight into the historical activity of a blockchain. I will examine a few metric and critical aspects connected to pricing to illustrate the parallels I observed between these on-chain metrics and price that led us to investigate this relationship further.

I picked some metrics including ones depict supply side (active supply, hash rate, network difficulty, gas fee.etc) and demand side (Active account, Total value locked, etc.). Different chain got different character, for example, ETH is the only one with POW+POS features, so it has metrics on both BTC and other smart contract chain. All the metrics can be found in the data related to this paper.

In general, on-chain analysis approach sheds light on the number of factors we should take into account when we are trying to model the dynamic movement of crypto currencies prices. In the next section, I will give a detail description on how those on-chain metrics help to build our research models.

3.1.3 Method to generate prediction

After I find out which input from on-chain data is useful for the input, I can use them as input, along with the spot price data (real-time price) to generate a more accurate prediction model for spot price of ETH. And at the same time we can calculate the lead and lag relationship between BTC and ETH, figure out after how many minutes is ETH going to change following BTC. Using this kind of information we can use time series method to predict the first-difference of ETH in the following time and combine spot price prediction and first-difference prediction we can get the predicted price after lag time of ETH.

3.2 Data used for research

To perform the research, I uses minute-level price data for BTC, ETH. The date range is from 1/1/2021 to 12/31/2021. All the data come from Binance, a well know crypto exchange.

But for the on chain data I did not got minute level. And as a matter of fact most people only use them as a reference for data analysis. The best we can

get is daily data. To form the data set, we only choose those average variables during 24hrs, which means the data during 24hr stays the same while price data changes every minutes. I hope this kind of method could enrich the input information for the model I build. All the data comes from Messari.com and Etherscan. All the data can be found in attached file.

As we perform time series analysis for time lag and perform LTSM model for ETH price predicting, we use minute level price data of ETH.

4 Modeling Process

4.1 Multivariate Linear Regression Model

4.1.1 Goal

We aim to build an intuitive and less computational-intensive model to quantify a linear relationship between BTC and ETH. During the process, we would also like to perform variable selection to an array of different on-chain metrics. Metrics that are selected in this process will be fed into the neural network models built later.

4.1.2 Model Preparation

Multivariate Linear regression has multiple independent variables contributing to the dependent variable and hence multiple coefficients to determine and complex computation due to the added variables. The equation of multivariate linear regression follows:

$$Y_i = \alpha + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + ... + \beta_n x_i^{(n)}$$

In our project, we have collected on-chain metrics for ETH and BTC on a daily level and here is 1 multivariate linear regression models that we need to conduct. Additional models can be found in the code.

Table 1: BTC-ETH Multivariate Linear Regression Output(Daily Data)

Variable	**Coef. for Price_Diff** ≤ 0.4	**Coef. for Price_Diff** ≥ 0.4
const	**-2.011**	**-0.4122**
CirculationSupply_Diff	**-4.4317**	**-0.8881**
MinerSupply_Diff	**-0.3099**	**0.0718**
Difficulty_Diff	**2.312**	**1.0434**
AddressCount_Diff	**3.5716**	**-0.0281**
TransactionFree_Diff	**2.455**	**0.4814**
R_squared	**0.543**	**0.416**
Adjusted_R_Squared	**0.535**	**0.408**
RMSE	**0.42386**	**N/A**

4.1.3 Variable Analysis

Generally, when it comes to multivariate linear regression, the best practice is not to throw in all the independent variables at a time and start minimizing the error function. First we should focus on selecting the best possible independent variables that contribute well to the dependent variable. For this, we go on and construct a correlation matrix for all the independent variables and the dependent variable from the observed data. We choose two correlation methods - Pearson, which is the commonly used method for financial markets because of the linear nature of the data, and Spearman's, a rank-based correlation method, to introduce new metrics potentially.

To illustrate the application of the two correlation methods, we have attached the results from the two methods applied on the BTC-ETH multivariate linear regression model. Figure 4 and Figure 5 show the correlation matrix results

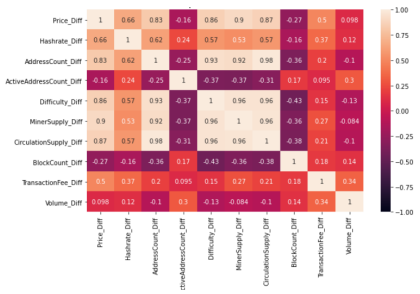


Figure 4: Spearman Correlation Results for BTC-ETH Multivariate Linear Regression

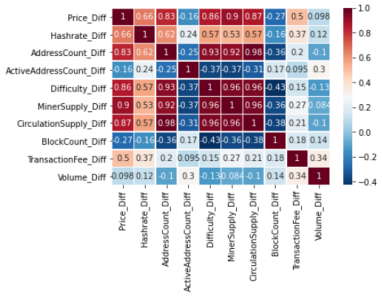


Figure 5: Pearson Correlation Results for BTC-ETH Multivariate Linear Regression

In order to have a more rigorous variable selection process, we have also used forward selection method by setting an alpha-to-entry value for including a new variable to the linear regression model to be 0.5. By repetitively running partial-F test, we first got a set of variables to conduct the multivariate linear regression models.

4.1.4 Regression Results

After running multivariate linear regression models on the data, we found in all cases we need to subdivide the data into two parts on the magnitude of the dependent variable. The reason for this division is due to the fact that the distribution for the dependent variable is bi-model, which violates the fundamental assumptions set for linear regression models. By subdividing the data based on a threshold level , we are tackling two sub-population separately, and each of the sub-data-set has normal distribution. Figure 6 displays one example of us doing the division of dependent variables on the BTC-ETH model

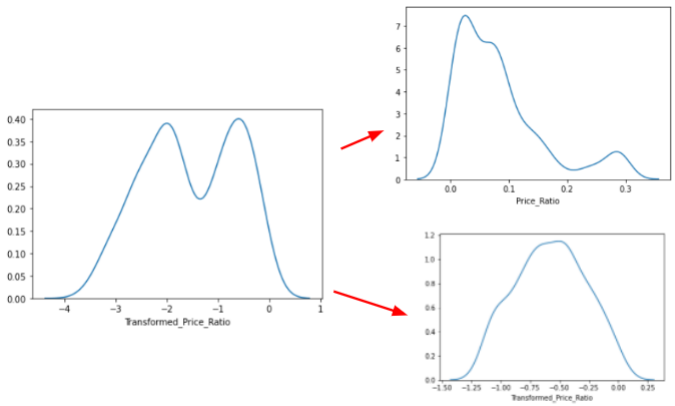


Figure 6: Subdividing the distribution of the dependent variable on BTC-ETH model

4.1.5 Problem showed up

Although we can get a relatively good expression on daily level data, but when we match average daily data with all the minute level price data and create a new dataset, we found the model can hardly explain the relationship between price and other metrics. The Adjusted R square dropped to 0.1 to 0.2 which means this kind of relationship cannot be explained in further research. We were trying to using lasso regression to narrow down the variables we would use but at last I have to give this idea up and include more variables in upcoming neural network model. One possible explanation is that although the price data can dramatically varies during a day, the onchain data we are using stays the same for a relatively long period. Although there is some of the variables can be useful like Total locked value and Gas Fee burned, this kind of variables will also increase the difficulty for the future model in neural network. We may better give the on-chain data up for now.

The key is to get the real-time(at least minute level data) for this analysis process. If possible in the future that we got the data in higher resolution, I think I can add this process again.

4.2 (Multi) Time Series Model

4.2.1 Goal

To better predict the price-movement relationship between trading coin pairs with Bitcoin (BTC) as the base currency, we developed a Vector Autoregressive (VAR) model to determine if a lead-lag relationship exists. Our hypothesis is that a lead relationship exists with BTC movement being a harbinger for price fluctuations in others.

Our original intention was to train the models on minute-level data to simu-

late real-world scenarios of model use. Unfortunately, we were not able to locate proper data that contained consistent minute-level data points in sequential order. We replaced with daily data under the assumption we can extrapolate our findings to minute-level data. Our analysis focused on the cross-covariance of Bitcoin (BTC), Ethereum (ETH).

4.2.2 Model Preparation

For our initial data exploration we plotted the Auto-covariance and Partial Auto-covariance functions for each coin individually. As expected, clear non-stationarity was present in each coin with little to concerning elevated spikes in the PACF indicating low levels of serial correlation. The full collection of ACF and PACF plots can be found in the Appendix.

Next, we merged the time series into trading pairs with BTC as the base running the ACF and PACF plots again. Non-stationarity was still present as expected with the emergence of consistent elevation in the PACF plot. The serial correlation plot with BTC as the lead displays numerous spikes outside the confidence bands. See appendix for additional cross-covariance plots.

To address the non-stationarity and serial correlation, we transformed the data by taking the first order difference of the log daily price. As shown in figure 7 below, the auto-covariance function (ACF) exhibits minimal spikes outside the confidence band indicating the series is now stationary.

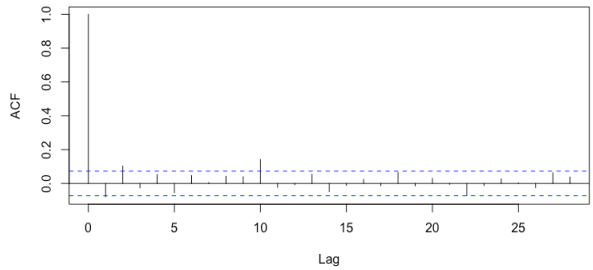


Figure 7: Auto-covariance Function for ETH and BTC

4.2.3 Order Selection and Fitting

We began the order selection by fitting each trading coin pair to the Vector Autoregressive (VAR) model for both AIC and BIC orders. We then applied a Wald Test of all the coefficients in each respective model to determine if whether we reject null hypothesis that the model performs similarly at lower orders. All of the models with the exception of ETH—BTC computed a p=value less than 0.05 signaling that we can reject the null hypothesis. The higher AIC order was selected for all models to maintain consistency. Next the models were transformed to restricted VARs to eliminate unnecessary variables that

provided marginal predictive power. The intention of restricting the models was for simplicity and computational efficiency. There was one exception to the additional fit of restricted VAR for XRP. Ther VAR order selection was optimized at 1 leaving no regressors to restrict so the model was only fitted unrestricted with p order = 1.

4.2.4 Forecasting and Model Performance

Our final test of the models was the lead-lag coefficients. This was performed again using a Wald Test to determine if Granger Causality existed. Our primary goal was to determine if the price movement of BTC could be used to forecast the price movement of another coin n periods ahead. Based on the p-values obtained in the test results we can conclude BTC does cause Granger Causality.

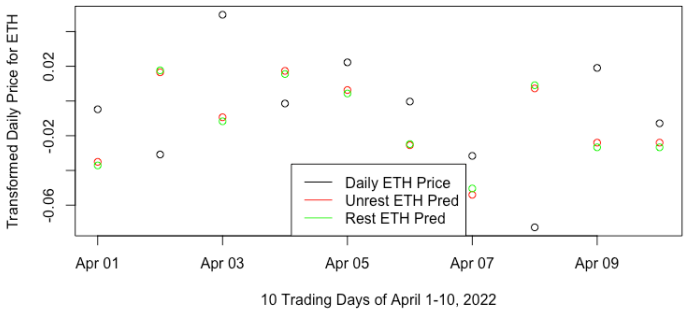


Figure 8: Unrestricted and Restricted VAR Predictions for April 1-10, 2022

4.2.5 Bring the Model to Minute Level

As we have already testified that there is such lead or lag relation exist, we need to know exactly what they are. There is method in multi time series called cross correlation which could measure it. Cross correlation is more useful when it comes to compare percentage of change, so we transformed price data to first-difference data. The visualiztion is shown in figure 9.

We used cross correlation checked the lag time and surprisingly there was initially no clear pattern. ETH lags BTC by 1minute but the relation is week. I looked back to the difference plot and found out as I was calculating half a million data, but most of the data stays in the normal range. If Bitcoin falls quickly or rises quickly, the FOMO would spread and there is always a following pattern. That might be what we need.

So we changed all the first-difference data in the range of (0.05, 0.95) percentile to 0, which means we don't consider the relatively common price movement's influence. The data left out are those moment with quick surge or fall in price which might be the market opportunities we are seeking. We performed cross correlation again and we do get a clearer lead-lag relation(As you can tell

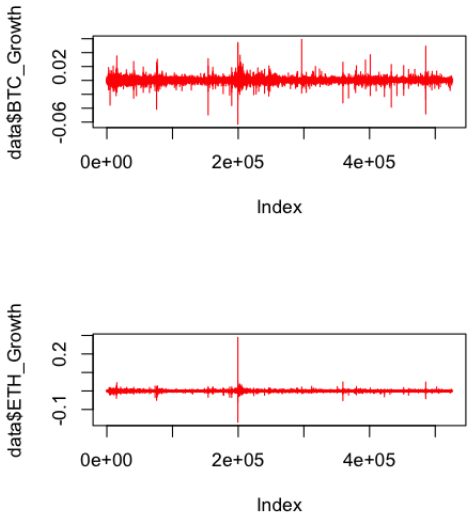


Figure 9: The first-difference of BTC and ETH

in figure 10, the upper graph is for regular first difference data, the latter one is for modified first difference data). As we can see after left out those common data the pattern seems stronger and finally we can say BTC lead ETH price by 1 minute.

Then we came to a point that the model frame need to change. Even though we got the lag relationship now but it came with condition. The lag only shows up when the price change is extremely big. So we only need to forecast the growth rate when there is big change in price. As for other situation we can simply predict by neural network.

Here we take BTC and ETH modified growth rate data and build a forecasting model by VAR model. This could be useful in the future when we calculate the finalized predicted price. The model and implementation can be seen in the code file.

4.3 Neural Network Model-LTSM Model

4.3.1 Goal

From analysis above we know that for the most of time the lead-lag relation won't show up when there is not big move in the market, we should better just perform a regular RNN-LTSM model to predict the price of the asset. As stated earlier, we found out that although on chain information would be a good input for daily price data, it has been absolutely different in minute-level data. Plus we don't have real minute level data (we are using average daily data and give

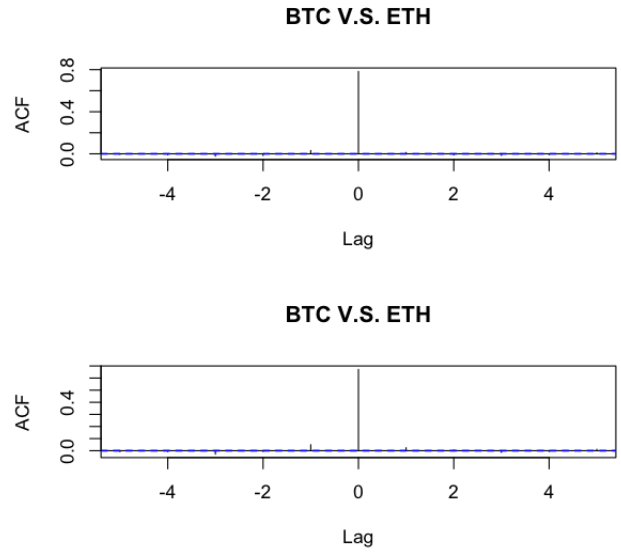


Figure 10: The first-difference of BTC and ETH before and after modify

it to every minute), so I at the end give up adding on-chain information for Neural Network Model Prediction. If we can afford professional real time data from those expensive blockchain data providers, we can try it in the future.

4.3.2 Model Preparation

An LSTM is another type of neural network, here TensorFlow with Keras is used as the wrapper for the model. The data is read into the model and then a Robust scaler is applied to the 'close price' and the rest of the dataset separately so that the 'close price' can quickly be inversed for graphs later. Functions are created to run the model, a split sequence function which will use the number of steps into the future to predict and the number of steps to learn from to produce windows of the data. A visualization function is created to visualize the training of the data over each epoch, visualisations of the loss and accuracy of the training and testing data, provide insight into the models performance and its ability to predict new data through the epochs.

A layer maker function is also created, the number of layers can easily be adjusted, by calling this function, the inputs required are the number of layers that are being called to add, the number of nodes in the layers, the activation function to be used and the dropout rate.

A validator function is created to create prediction values for every interval, this will then be used to assist in creating a future prediction for the currencies. A validation mean-squared-error function is created, to calculate the MSE between the prediction and actual data frames.

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 60, 90)	34560
lstm_1 (LSTM)	(None, 60, 30)	14520
lstm_2 (LSTM)	(None, 60, 30)	7320
lstm_3 (LSTM)	(None, 60)	21840
dense (Dense)	(None, 1)	61

Total params: 78,301
Trainable params: 78,301
Non-trainable params: 0

Figure 11: LTSM Design

4.3.3 Model Outcome

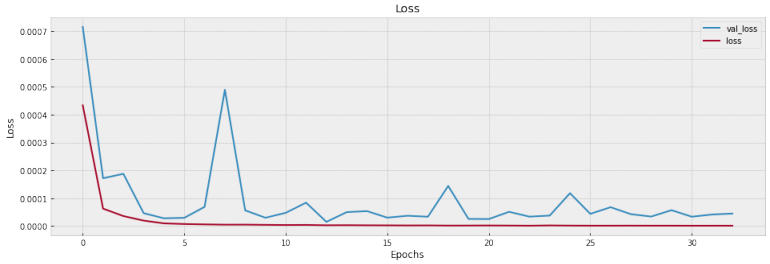


Figure 12: Loss Measured from model

5 Conclusion

We started from an ideal assumption and framed a good strategy at the beginning of this project. We thought it could be simple and easy to compute but at last we found out that the real life problem is much more complicated.

First, in minute level price data we can't match it with average daily metrics. So until the time we can access real time on chain data, we can't figure out if adding more variables to the model will help. At least for now it's not an efficient way to do it.

Second, the lead-lag relationship exist in the situation when the price of BTC changes really big in short time, that means our initial thought to build strategy on lead-lag relationship only fit in certain time. This also remind us to be patience when investing.

Third, in special situation, we can do can use the lead-lag relationship to predict the relevant growth rate at that time. As all these coins stay in 1min lag, we can calculate the coming price as RNN predicted price * VAR predicted

growth rate. This is our chance to trade. For the ordinary situation, we need further research in RNN model to predict.

Besides the dataset might also influence the result, we used to use a dataset from Bitfinex which shows a clearer lead-lag relation, but there are always some record missing so we changed to the current dataset. We may gather more data and give different try.

Restricted by page limitation, most of the report focused on BTC-ETH trading pair. More information and result can be found in our program.

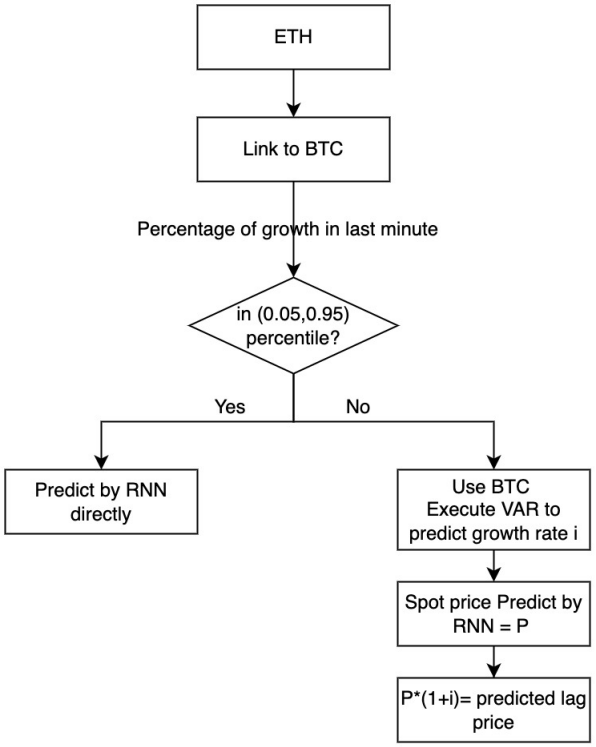


Figure 13: The result of current stage model, slightly different from the frame

6 References

1. Akbulaev, N., Mammadov, I., Hemdullayeva, M. (2020). Correlation and regression analysis of the relation between Ethereum Price and both its volume and Bitcoin Price.

2. Laura Alessandretti, Abeer ElBahrawy, Andrea Baronchelli. (2018) Anticipating Cryptocurrency Prices Using Machine Learning

3. Reaz Chowdhury, M. Arifur Rahman, M. Sohel Rahman, M.R.C. Mahdy (2019) An approach to predict and forecast the price of constituents and index of cryptocurrency using machine learning

4. Han-Min Kim, Gee-Woo Bock, Gunwoong Lee (2020), Predicting Ethereum prices with machine learning based on Blockchain information

5. Andrew Burnie (2018), Exploring the Interconnectedness of Cryptocurrencies using Correlation Networks

6. Wang, Y., Zhou, L., Chen, Y., Shuiguang, D., Haotian, W., Wattenhofer, R. (2022). Cyclic Arbitrage in Decentralized Exchanges.

7. Santosha Kumar Mallick: Causal relationship between Crypto currencies: An Analytical Study between Bitcoin and Binance Coin

8. Yhlas Sovbetov: Factors Influencing Cryptocurrency Prices: Evidence from Bitcoin, Ethereum, Dash, Bitcoin, and Monero