Project Proposal: Cryptocurrency Price Movements and Machine Learning

Runhua Li, Xinyu Liu, Joanna Zhang, Ying Zhou

Jan 24, 2020

1 Introduction

Our project aims at forecasting cryptocurrency future prices based on past prices. Nowadays, machine learning has become a handy tool in explaining trends in the financial markets, such as asset price movements. Literature in applying machine learning algorithms in both the equity markets [HNW04, SAF15] and the cryptocurrency markets [AEAB18, RKDP18] has recognized machine learning as a reliable instrument. We will use Long Short Time Memory (LSTM) to deal with time-series data. If Possible, we may also explore the non-classical and opinion-based data, such as text information from Twitter, using sentiment score tool to generate potentially useful supplement information for market prediction.

The motivation of this project is to test the weak form of market efficiency in cryptocurrency markets. The weak form of market efficiency states that future stock prices are not influenced by past events, and thus analyzing past prices is not helpful for price prediction. We intend to investigate whether it is actually advantageous to analyze cryptocurrency past prices in the crypto market.

2 Relevant Literature

Pioneer research papers on using machine learning as a forecasting tool primarily come from studies focusing on the equity market. Huang et al. [HNW04] compare a model based on Support Vector Machine (SVM) and the random walk model in predicting the movement direction of the NIKKEI 225 index. The paper concludes that the SVM model makes more accurate predictions compared to the traditional random walk model. Enke and Thawornwong [ET05] and Sheta et al. [SAF15] both demonstrate that the strategies guided by neural network classification models generate higher profits under the same risk exposure than the traditional linear regression models.

In the area of crypto-related research, existing literature has explored many different ways of forecasting cryptocurrency price movements. Alessandretti et al. [AEAB18] build their cryptocurrency forecasting algorithms based on XGBoost and compare the algorithms with a baseline strategy of taking the moving average of prices in the past few days. The paper attempts to build a crypto portfolio that is based on the predictions. The portfolio outperformed the baseline strategy on a daily basis and during the whole period considered in the paper. Rebane et al. [RKDP18] apply a recurrent neural network (RNN) approach to analyze Bitcoin time series price data. The results confirm that the RNN method may improve the classical autoregressive integrated moving average (ARIMA) model since RNN generates more accurate predictions.

In addition to past prices, Google trend is also considered as an important indicator of cryptocurrency prices. Hu et al. [HTZW18] find that using a neural network model that takes Google trend into consideration can achieve a hit ratio of 88% in predicting stock prices. Researchers in academia have also considered combining sentiment analysis with machine learning in the field of cryptocurrency. Colianni et al.[CRS15] employ a third-party open-source sentiment analysis API to and several machine learning models to examine the relationship between twitter sentiment towards Bitcoin and Bitcoin price movements. Their research finds that the Bernoulli Naive Bayes model can achieve a day-to-day accuracy of 95%, while the SVM model reached a day-to-day accuracy of 83%.

3 Data

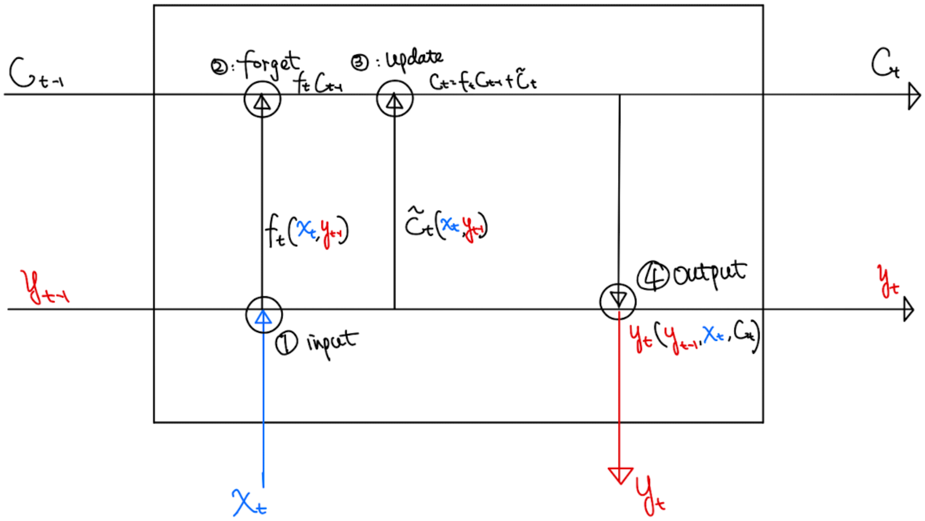
From a time series machine learning approach, our project will primarily rely on historical daily information including price, trading volume, market cap, and the option prices of some major cryptocurrencies such as Bitcoin and Ethereum. As of now, we have constructed a preliminary database to support our further experiment on modeling. Our cryptocurrency data comes from coinmarketcap.com, a leading cryptocurrency market information provider.

To test whether it is useful to include factors other than prices into the model, we obtain the data on bitcoin's Google trend from Google. Google trend is an index reflecting the number of searches that is calculated and save on a daily base. For sentiment analysis, we will rely on twitter's Developer API to filter relevant twitters as the input of our Natural Language Processing (NLP) based sentiment package to generate sentiment scores.

4 Methodology

For time series prediction, we will be first testing a specific type of recurrent neural network called the Long Short Time Memory (LSTM). The reason why we attempt to use LSTM to predict cryptocurrency prices is its strength in handling ”long-term dependence” problem. That means, when the prediction output depends on not only recent inputs, but also distant inputs, LSTM is good at utilizing information contained in distant past inputs. If there is ”long-term dependence” problem, then LSTM can excel ordinary RNN in prediction.

Generally, LSTMs can utilize distant past inputs' information to make predictions because of its 4-step procedure to process both past outputs and current input, i.e., a 4-layer structure suggested by [Ola15]. The four steps are to ”forget”, ”input”, ”update”, ”output”. See below a hand-drawn illustration.

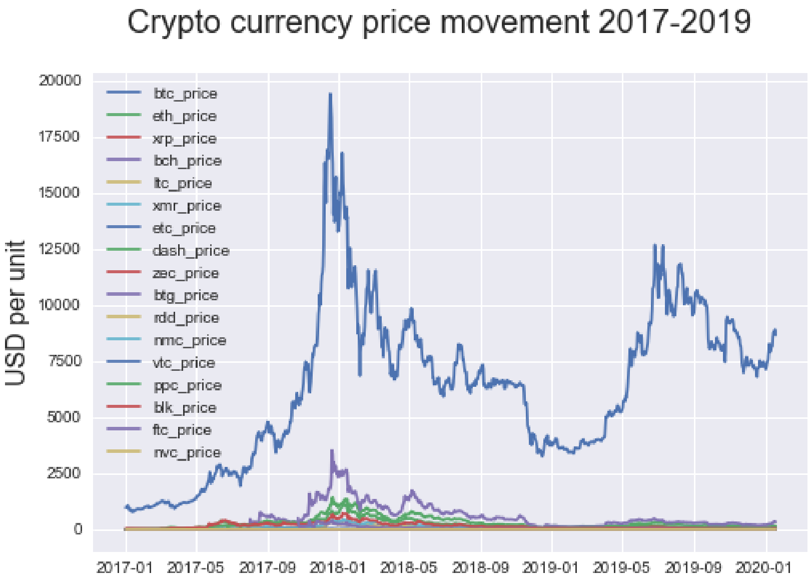


LSTM is constructed by a series of boxes aligned horizontally in a time-series manner. The top line in the box is called cell, a vector that keeps useful inputs. First, The previous output y_{t-1} and input x_t are used in the first step to generate a vector f_t that dictates the forgetting process. Second, the previous cell vector is modified according to f_t. Third, the modified cell plus an update vector $\tilde{C}_t$, which is generated from the input and the previous output, became the updated cell. Finally, an output is generated according to the updated cell, the current input and the previous output.

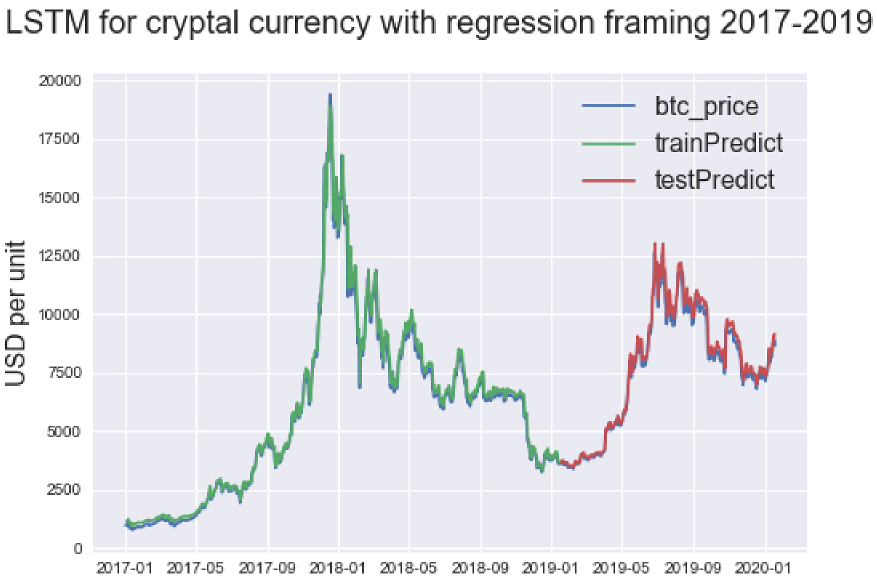
In our project, the cell can contain useful past price data, which are examined and modified each time a new data is observed. We then make predictions using both the cell and the new input data. The four steps can involve transformations using sigmoid function and tanh function.

5 Preliminary Progress

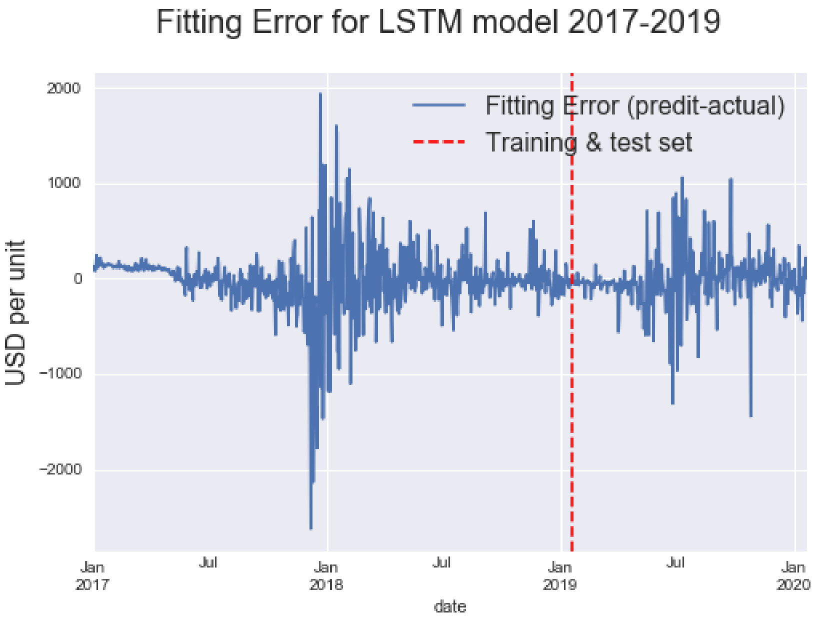
Using the past prices of Bitcoin that we have already obtained, we are able to present the price movements between 2017-2019 in the graph below. It is clear that there exists an autoregressive pattern in the price movements as well as an autocorrelation pattern in price volatility, which suggests that we can try to develop a learning model that captures these nonlinear effects. From the graph below we can tell that compared to other currencies, Bitcoin prices are much higher than other coins. Since Bitcoin is a dominant coin in the cryptocurrency markets, therefore it can be used to do a preliminary test on our hypothesis.



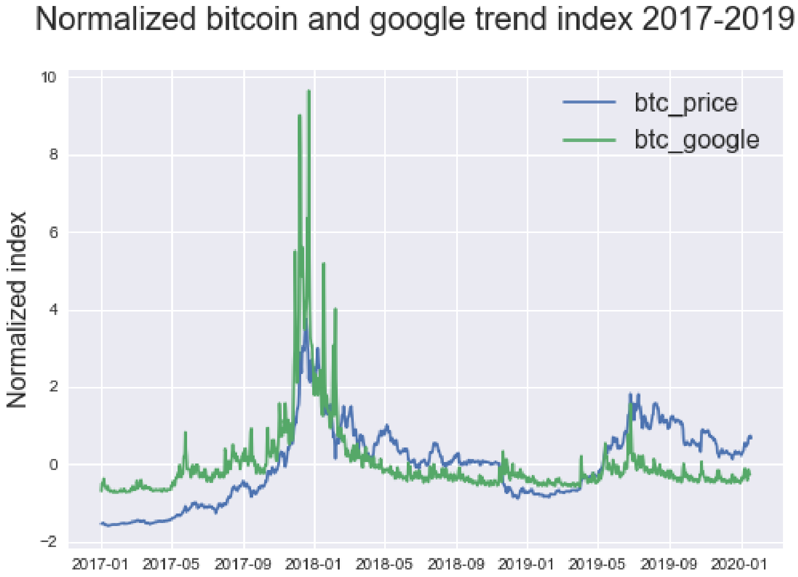
With the LSTM model described in the methodology section, we implemented the model on Bitcoin time-series price data. With 100 times of iteration, we reached a train score of 375.74 RMSE, and a test score of 359.23 RMSE.



In addition, we briefly examined the fitting error from the LSTM model, from which we noticed that our current model seems to be weak in terms of timely capturing price movement, and errors tend to cluster. This will be further improved in the future.



Lastly, we tested the possibility to include predictors other than past prices into our model. Here we plot the mean standard deviation normalized daily price of Bitcoin against its global Google trend. The Google trend index reflects the number of searches with keywords including ”Bitcoin”. Unsurprisingly, the picture demonstrates a strong correlation (with a Pearson correlation coefficient of 58.4%) between the number of searches and Bitcoin prices. This implies that other non-price factors related to Bitcoin may also be used as a predictor.



References

[AEAB18] Laura Alessandretti, Abeer ElBahrawy, Luca Maria Aiello, and Andrea Baronchelli. Anticipating cryptocurrency prices using machine learning. *Complexity*, 2018.

[CRS15] Stuart Colianni, Stephanie Rosales, and Michael Signorotti. Algorithmic trading of cryptocurrency based on twitter sentiment analysis. *CS229 Project*, 2015.

[ET05] David Enke and Suraphan Thawornwong. The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications*, 2005.

[HNW04] Wei Huang, Yoshiteru Nakamoria, and Shou-Yang Wang. Forecasting stock market movement direction with support vector machine. *Computers Operations Research*, 2004.

[HTZW18] Hongping Hu, Li Tang, Shuhua Zhang, and Haiyan Wang. Predicting the direction of stock markets using optimized neural networks with google trends. *Science Direct*, 2018.

[Ola15] Christopher Olah. *Understanding LSTM Networks*, 2015. https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[RKDP18] Jonathan Rebane, Isak Karlsson, Stojan Denic, and Panagiotis Papapetrou. Seq2seq rnns and arima models for cryptocurrency prediction: A comparative study. *SIGKDD Fintech*, 2018.

[SAF15] Alaa F. Sheta, Sara Elsir M. Ahmed, and Hossam Faris. A comparison between regression, artificial neural networks and support vector machines for predicting stock market index. *Soft Computing*, 2015.