Bitcoin Price Prediction using Machine Learning

Z. Ez-zahar*, A. Al Horami*, H. Hrimech*

*Department of Computer Engineering and Mathematics - National School of Applied Sciences – Hassan I University – Morocco

ez-zahar.ensa@uhp.ac.ma, alhorami.ensa@uhp.ac.ma, hrimech@hotmail.fr

Abstract :

The foundation for peer-to-peer financial transactions based on blockchain technology is provided by decentralized cryptocurrencies, such as bitcoin, which is a class of digital asset. Price volatility is one of the key issues with decentralized crypto currencies, which emphasizes the importance of understanding the underlying price model. Additionally, the statistical distribution of data for Bitcoin prices exhibits nonstationary behaviour over time. In order to forecast Bitcoin price fluctuations and pricing in the short and medium term. This work aims to study high-performance machine learning-based classification and regression algorithms.

Keywords: Machine learning, Blockchain, Bitcoin price prediction, regression algorithms, crypto currency.

1. INTRODUCTION

The most significant disruption to economies and financial institutions currently occurring is the result of digital transformation. At a previously unheard-of rate, the global financial and economic systems are going digital. A recent analysis estimates that the size of the digital economy in 2025 will be 25% (or USD 23 trillion), made up of both tangible and intangible digital assets [1]. Distributed ledger technology (DLT), whose most well-known use is the cryptocurrency known as Bitcoin [2], is the most modern technology for creating and using digital assets. As a result of these developments, blockchain technology has established itself at the nexus of Fintech and next-generation networks [3].

Price volatility is a significant problem with non-tangible digital assets, particularly cryptocurrencies. Figure 1 displays the transaction volume from April 1, 2013, to December 31, 2018. BTC values have fluctuated wildly during this time. BTC had low values, limited use in virtual transactions, and low public interest prior to 2013. Our models do not account for that period. BTC is a digital asset that is highly durable and can restore value after major decreases in price, especially when market uncertainty is severe, like it was during the COVID-19 epidemic [4]. Despite the BTC pricing' extreme volatility.

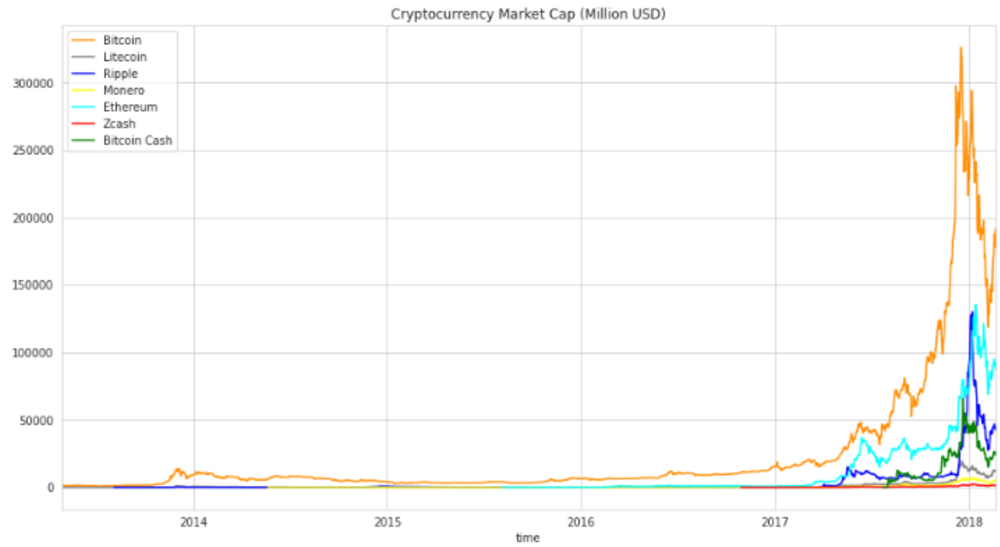

Figure 1. Cryptocurrency Market Cap (Million USD)

Considering the magnitude of the BTC price movements that have place after these dates, a recent analysis is required. Second, the stated studies concentrate on price increase/decrease forecasting for the following day's pricing as well as end-of-day closing price forecasting. In this study, the daily closing price forecasts, mid-term price increase/decrease forecasting, and short-term price increase/decrease forecasting (for the end of the day and the following day) are discussed. The forecasting horizons range from 7 days to 30 days. Additionally, this is the first study to use machine learning to anticipate end-of-day, short-term (7 days) and long-term (30 days) BTC prices while considering all price indications through December 31, 2018.

Our performance statistics show that the most recent research in terms of daily closing price forecasting and price increase/decrease forecasting are outperformed. For medium-term (7, 30 day) BTC price projections as well as price increase/decrease forecasting, also, the high-performance neural network-based models are also presented in this work.

2. ONGOING WORK AND ACHIEVED RESULTS

Creating a database is the first step in Bitcoin prediction. To study our project, data is collected from the following sources:

A. Web scraper:

Features and price information for BTC are readily accessible online. Using a Python 3.6 web scraper, the data for this study were obtained from https://bitinfocharts.com. Based on technical indications, more than 700 features were gathered. A smaller subset of features was chosen using the feature selection approach from this vast feature set. Simple Moving Average (SMA), Exponential Moving Average (EMA), Relative Strength Index (RSI), Weighted Moving Average (WMA), Standard Deviation (STD), Variance (VAR), Triple Moving Exponential (TRIX), and Rate of Change are the technical indicators (ROC).

B. Kaggle:

All historical data for each cryptocurrency's open, high, low, close, trading volume, and market capitalization.

The daily close rate, min-maxed with the high and low values for the day, is known as the close ratio.

Spread is the $USD difference between the high and low values for the day.

Close Ratio is calculated as (Close-Low)/(High-Low).

Database normalization is the following stage. This stage is done to achieve consistency, i.e., minimize or get rid of redundant data, unimportant points, and other redundancies. Five different normalization methods are applied to our data. The results obtained after implementing various normalization techniques are presented below:

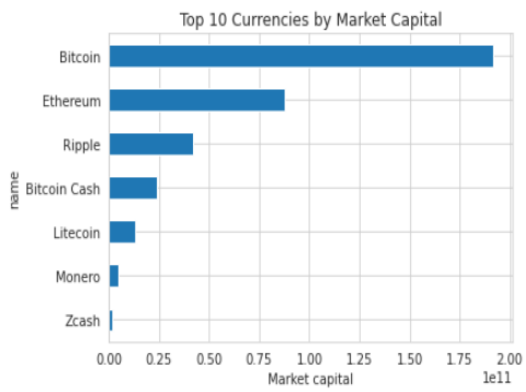

Figure 2. Top 10 Currencies by Market Capital

Figure 2 shows the top 10 currencies by Market Capital. Compared to other currencies, Bitcoin has a very significant market capital. Only these 10 cryptocurrencies are included to make visualization easier because they are the first 10 currencies and because our data has more than 9 lakh lines.

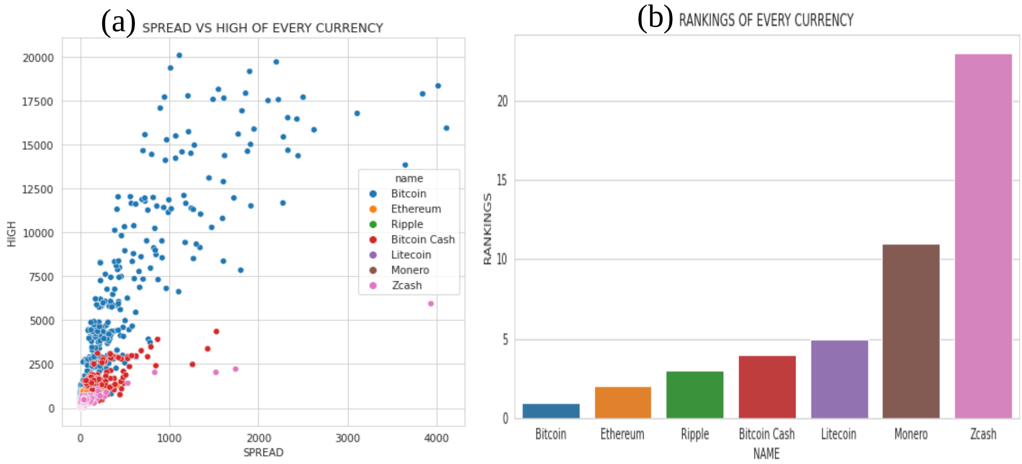

Figure 3. a) Spread vs high of every currency, b) Rankings of every currency.

In relation to the spread of each currency, the point nuage shows the highest value (see Figure 3). The graph demonstrates that all currencies have extremely low spreads and highs, except for bitcoin. It can be concluded that Bitcoin is the currency with the greatest market capitalization and highest perceived value.

Figure 4 shows that is nothing much happened before 2017 (the exception being a few transactions for Bitcoin). From the volume of transaction graph, 2017 appears to have been a year of disruption for the cryptocurrency industry. In addition, the amount of transactions for other cryptocurrencies increased in proportion to the growth of BitCoin's volume, and the cost of Bitcoin has increased much like the cost of other cryptocurrencies, also it appears that the market changes are being led by the Bitcoin cryptocurrency.

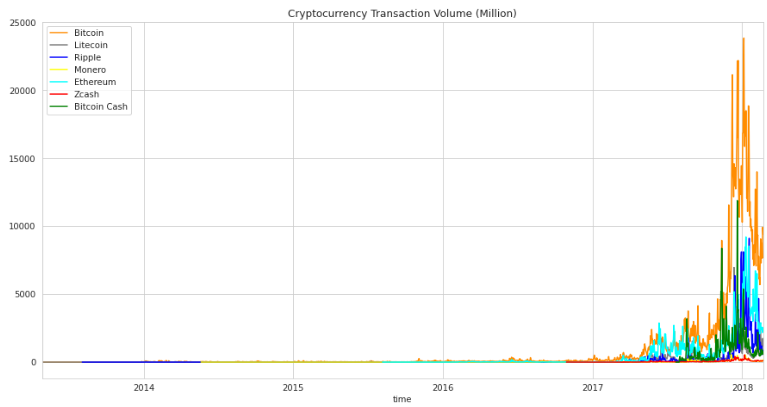

Figure 4. Cryptocurrency Transaction Volume (Million)

Cryptocurrency transaction volume is presented in Figure 5, it is clear that the volume of BitCoin increased and peaked around the 12/7, other crypto-currencies started to increase and peaked a few days later by closely examining the monthly data for transaction volume in December 2017.

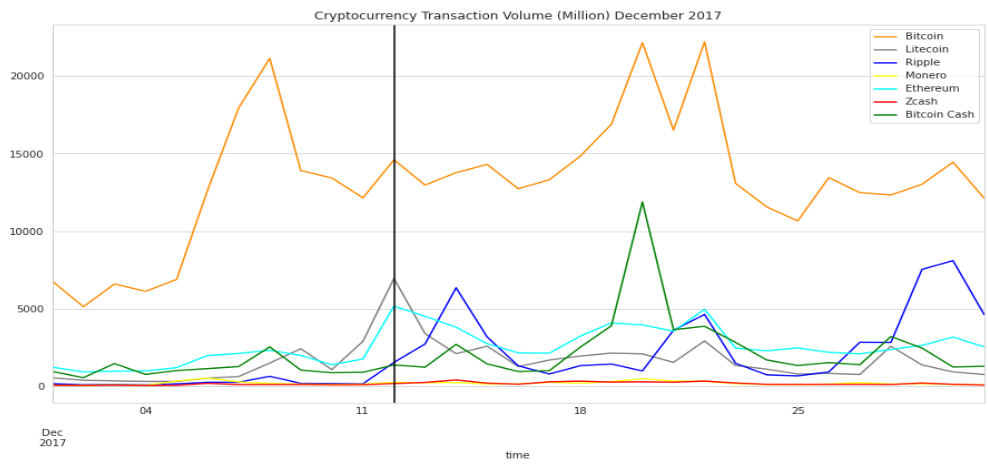

Figure 5. Cryptocurrency Transaction Volume (Million) December 2017

3. Methodology

An essential subfield of artificial intelligence is machine learning (AI). It can be separated into supervised learning, unsupervised learning, and reinforcement learning depending on whether a goal variable is present. This study uses a regression function with supervised learning to forecast future Bitcoin prices. An algorithm is programmed, a learner is formed, and a high-precision learner is obtained by repeatedly training the learner using training data and the process of validation. This is the unified execution logic of machine learning. In order to evaluate and apply the test results, the trained learner is finally substituted.

The open-source machine learning package for Python is used in this study to implement both random forest regression and LSTM model training. Sklearn is the library used by random forest regression, and keras is the research library for LSTM. Pandas handle the pre-processing and data collection.

A. Random Forest:

An ensemble of several regression trees is called a random forest. High explicability is a benefit, but the anticipated results are constrained by the training samples. The classification is based on making the average of the sum of squared residuals of each group the least, as indicated in Equation (1) below. The regression tree's basic idea is to split the parent group into subgroups using an indicator of a specific variable.

$$\frac{1}{n_1}\sum_{i=1}^{n_1}\left(y_i-\overline{y_{(1:n_1)}}\right)+\frac{1}{n_2-n_1}\sum_{j=n_1+1}^{n_2}\left(y_j-\overline{y_{(n_1+1:n_2)}}\right)\rightarrow min \qquad \text{Equation (1)}$$

B. LSTM :

The RNN algorithm is distinct from the traditional DNN algorithm. When data is inserted into the model, it will not only provide an output value but also change the model's parameters. The RNN algorithm can keep track of prior input data information in the model. In order to correct RNN's short memory flaw, the LSTM model is used in this study. Through the routes of the three activation functions of Forget Gate, Input Gate, and Output Gate, data modifications are made to the RNN model and the memory model.

The LSTM model layout for this experiment is as follows based on the property that the output value of the LSTM model can be resubstituted into another layer of the LSTM model and the use of the dropout layer indicated in the literature. I tested [min = 10%, max = 50%] for each dropout layer when it came to parameter setup. It turns out that there is a phenomenon known as overlearning when the training data performs well but the prediction error of the validation data is high when the overall value of dropout is minimal. The mistakes of the training data and the validation data are both significant when the total value of dropout is set to a high value. The investigation also revealed that descending order has worse dropout value prediction accuracy than ascending order. Each layer's activation function is set to "linear," which performs better than "sigmoid" and "tanh." Table 1 and Figure 6 below illustrate how the LSTM's special value and framework work. Validation data are the final 5% of the training data.

Table 1. Details of the LSTM model.

Model: "sequential"

Layer (type)	Output Shape	Param #
bidirectional (Bidirectiona l)	(None, 49, 98)	34104
dropout (Dropout)	(None, 49, 98)	0
bidirectional_1 (Bidirectio nal)	(None, 49, 196)	154448
dropout_1 (Dropout)	(None, 49, 196)	0
bidirectional_2 (Bidirectio nal)	(None, 98)	96432
dense (Dense)	(None, 1)	99
activation (Activation)	(None, 1)	0

Total params: 285,083
Trainable params: 285,083
Non-trainable params: 0

None

After applying algorithms, the results obtained are presented below.

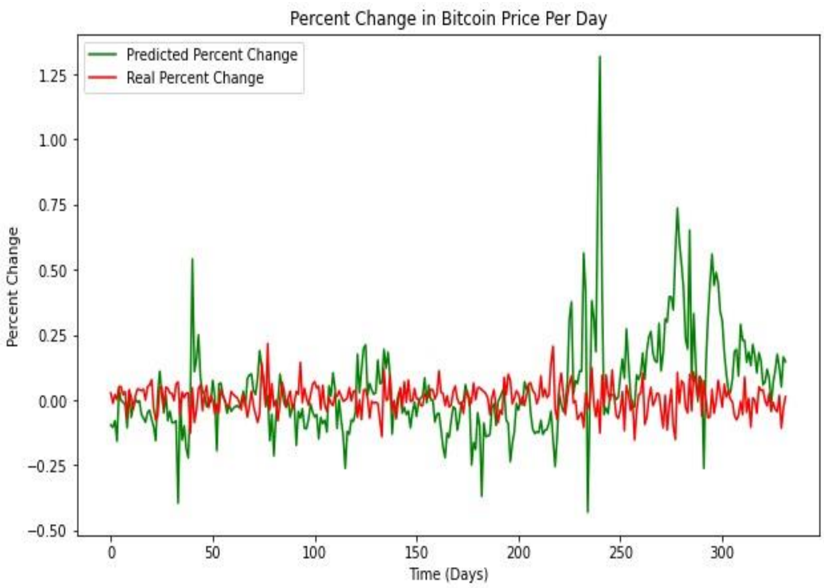

Figure 6. Percent change in Bitcoin Price per Day.

Considering the expected short-term trend for the coming month (see Figure 6), the price of bitcoin is expected to fluctuate further. A downward trend is immediately followed by an upward trend in prices, primarily due to a lack of vigor. Investors need to exercise caution given the price's dramatic fluctuations. It is advised that traders purchase both a short-term sale option and a long-term purchase option.

4. CONCLUSIONS

We plan to apply the approaches stated above to solve the Bitcoin prediction problem after creating the learning framework and completing the normalization.

ACKNOWLEDGMENT

We would like to express our gratitude to our parents for their unwavering support as well as to our project mentor, Mr. Hamid HRIMECH, for inspiring us to work on this project.

REFERENCES

[1] Xu W, Cooper A (2017) Digital spillover: measuring the true impact of the digital economy. Huawei and Oxford Economics. https://www.huawei.com/minisite/gci/en/digital-spillover/files/ gci-digital-spillover.pdf. Accessed 2 July 2020

[2] Oliver J (2013) Mastering blockchain distributed ledger technology, and smart contracts explained-packt publishing, vol 53. Packt Publishing Ltd, https://doi.org/10.1017/ CBO9781107415324.004, arXiv:1011.1669v3

[3] Unal D, Hammoudeh M, Kiraz MS (2020) Policy specification and verification for blockchain and smart contracts in 5G networks. ICT Express 6(1):43–47. https://doi.org/10.1016/j.icte.2019.07.002

[4] Shukla S, Dave S (2020) Bitcoin beats coronavirus blues. https:// economictimes.indiatimes.com/markets/stocks/news/bitcoinbeats-coronavirus-blues/articleshow/75049718.cms