2020 Management Science Informatization and Economic Innovation Development Conference (MSIEID) | 978-1-6654-1541-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/MSIEID52046.2020.00050

Bitcoin price prediction in the time of COVID-19

Jiayun Luo[1]
[1]Department of Statistics, University of California-Los Angeles, Los Angeles, 90024, USA

***ABSTRACT*: Based on the Bitcoin exchange data, COVID-19 data, and Twitter data from January 2020 to July 2020, this paper compares the performance of four different machine learning models on predicting the Bitcoin return rate and price trend. Data are formulated to four input feature sets, including: (1) Historical Bitcoin exchange data; (2) Historical Bitcoin exchange data + COVID-19 data (recovery, confirmed, death); (3) Historical Bitcoin exchange data + Twitter data; (4) Historical Bitcoin exchange data + COVID-19 data (recovery, confirmed, death) + Twitter data. The four machine learning models implemented are: (1) Random forest; (2) Decision tree; (3) AdaBoost; (4) Support vector machine. We found that: (1) Twitter data can improve the performance of models; (2) People consider information within 5 days when they make decisions on investments; (3) Support vector machine does not perform well in predicting Bitcoin return rate or price trend; (4) COVID-19 data does not help improve the prediction. However, we have very limited COVID-19 data, so future research with more COVID-19 data may help confirm if the last statement is correct or not.**

Keywords: Bitcoin, price prediction, COVID-19, Twitter, sentiment analysis

I. INTRODUCTION

Bitcoin is currently one of the most popular cryptocurrencies around the world. However, it is not the first cryptocurrency that appears in the world. Satoshi Nakamoto first introduced Bitcoin and the concept of Blockchain in one of his papers in 2008 [1]. After the paper, Bitcoin started to circulate among computer geeks. But only until consecutive increases in Bitcoin price that people start to notice it and Bitcoin becomes a hot topic. Nowadays, despite the appearance of many other cryptocurrencies, Bitcoin still ranks the top on cryptocurrency popularity.

As Bitcoin's price has very high volatility, many people are attracted to the research of predicting the price trend of Bitcoin because investors have a chance to gain high profit from the price change. As more and more researches focus on the topic, the methods used for prediction also expand from only time series models to machine learning models and deep learning models [2]. The input feature sets also expand from historical price and exchange volume to text data from social media, news, internet search, and more.

As financial equipment that is not closely related to real economics, Bitcoin is always used as an effective tool for hedging the risk of products directly affected by the macroeconomic market. However, in the year 2020, the explosion of the coronavirus pandemic has forced the world to change, created a large effect on real economics, and also indirectly affecting the financial market, including the cryptocurrency market. Therefore, we want to include this rare event in our model to see if there are any significant effects.

In this paper, we predict both the price trend and the return rate of Bitcoin based on four different machine learning models. Besides the comparison of four ML models' performance, we compared the significance of four different input feature sets, including (1) Historical Bitcoin exchange data; (2) Historical Bitcoin exchange data + COVID-19 data (recovery, confirmed, death); (3) Historical Bitcoin exchange data + Twitter data; (4) Historical Bitcoin exchange data + COVID-19 data (recovery, confirmed, death) + Twitter data. Finally, we consider the influence of lag of time and we have the following findings: (1) Twitter data can improve the performance of models. (2) People consider the information within 5 days when they make decisions on investments. (3) Support vector machine does not perform well in predicting Bitcoin return rate or price trend. (4) COVID-19 data does not help improve the prediction. However, we have very limited COVID-19 data so future research with more COVID-19 data may help confirm if this statement is correct.

In the following sections, we would explain more detail about our work. The second section of this paper would briefly summarize the previous work on this topic. The third section would introduce the dataset we used for research. The fourth section would explain the results we get from the various combinations of models and input feature sets. Finally, in the fifth section, we provide our conclusions and discuss the future direction for improvement.

II. RELATED WORK

A. Bitcoin price prediction

There is already many relevant research in the literature on the topic of Bitcoin price prediction. Here in this section, we briefly introduce some recent papers on this topic.

[3] used Google News data and Reddit messages data for sentiment analysis using Valence Aware Dictionary and sEntiment Reasonor (VADER), TextBlob, and Flair. Results from sentiment analysis and historical price of Bitcoin, Litecoin, and Ethereum are combined for training the machine learning models. (LSTM, GRU, 1D-CNN). All data were collected from January 1, 2018, to November 20, 2019, with a time interval of one hour, except Google News, which were collected per day basis. When using all the features, the LSTM model with the sum of all the sentiment analysis values gives the lowest test RMSE of 434.87. However, an interesting finding is that without adding the sentiment analysis data, the LSTM model gives a test RMSE of 116.36, which is the overall minimum.

For short-term cryptocurrency price prediction, [4] compared SVR, Stochastic Gradient Descent, Gradient Boosting Model, Multilayer Perceptron Neural Network, Least Squares Linear Regression, AdaBoost, Bayesian ridge regression, Decision tree, ElasticNet, and their Hybrid, i.e., mean of all models. Different cryptocurrencies are considered,

978-1-6654-1541-5/20/$31.00 ©2020 IEEE
DOI 10.1109/MSIEID52046.2020.00050

including Bitcoin (BTC), Ethereum (ETH), Electroneum (ETN), Ripple (XRP), ZEC Cash (ZEC), and Monero (XMR). Twitter sentiments and Google Trends data were used as input features. The authors conducted an empirical experiment by trading $100 on the BitBay cryptocurrency exchange over 1 month and the account balance stood at $114.82. In contrast, when they used KryptoBot, a well-known tool for cryptocurrency trading, they managed to convert $100 into $102.45 within the same period.

[5] collected data for five currencies in OTC, currency, and contract market, namely Bitcoin (BTC), Ethereum (ETH), Tether USD (USDT), EOS, Ripple (XRP), litecoin (LTC). Accordingly, it also crawled daily user reviews on online forums from the third quarter of 2018 to the first quarter of 2019. Specific data items include the content and quantity of title, comments, replies, and time of publication. They concluded that the LSTM model has the best performance, and the addition of comments' sentiment can significantly improve the accuracy of prediction. Moreover, users refer to trading data and comment data 1-7 days before when making trading decisions.

[6] tested the Bitcoin price predictability using intraday BTC–USD returns and investor sentiment and investor attention measured from StockTwits data. Additionally, they consider a potential reverse causality - returns affecting investors' sentiment - by conducting Granger causality tests. The result shows that the sentiment factor is statistically significant when they consider the lag of time of up to 15 minutes, and investor attention, which is measured by the total number of messages in a given time interval, is not significantly affecting Bitcoin return.

B. Impact of COVID-19 on Financial Market

Some studies have focused on the effect of COVID-19 on the stock market. [7] examined the impact of COVID-19 on emerging stock markets. It found the negative impact has gradually fallen and begun to taper off by mid-April. It also found the highest impact is in Asian and the lowest in European emerging markets. [8] examined the March 2020 stock market crash triggered by COVID-19. Different sectors show different influences. Natural gas, food, healthcare, and software stocks earn high positive returns. Petroleum, real estate, entertainment, and hospitality stocks fall dramatically. [9] also examined the stock return predictability during the COVID-19 crisis. A novel robust Lasso approach with Cauchy errors was adopted for predictive regressions. The results showed that corporate bonds, both investment grade and high yield, had significant predictive ability.

The cryptocurrency market during COVID-19 is also studied. The exploration of the cryptocurrencies market efficiency before and after the COVID-19 pandemic through a multifractal analysis was conducted in [10]. COVID-19 was revealed to have an impact on the efficiency of all the five cryptocurrencies. In [11], 36 statistical tests are performed to check for differences between periods of time (pre- versus during COVID-19 pandemic samples) on the one hand, as well as check for differences between markets (cryptocurrencies versus stocks), on the other hand. Cryptos showed more instability and more irregularity during the COVID-19 pandemic compared to international stock markets.

III. DATASET

In this research, we used historical Bitcoin transaction data, Twitter data, and COVID-19 data. We collected data from 1st January 2020 to 31st July 2020 from online open sources. Historical Bitcoin transaction data includes Open, High, Low, Close prices, and Trading Volume. We download this data from Yahoo Finance.

Twitter data is crawled by using a tool named "Twint" [12], because API from twitter has a rather strict limit on how much data you can collect each time. During crawling, we use the keyword "Bitcoin" and collect all the tweets with this keyword in either the content or hashtags. After collecting all the tweets, we remove those that are apparently ads or not relevant. Then we used the "nltk" package to perform sentiment analysis (specifically, Valence Aware Dictionary and sEntiment Reasonor (VADER)) on the content of tweets. We take the average of all the polarity scores for each day as the representation of sentiment of that day. Furthermore, we gather information including likes amount, replies amount, and retweets amount for all the tweets per day as future feature inputs.

According to [13], the top 7 countries with the most Bitcoin holders are United States, Romania, China, Spain, Japan, Switzerland, and South Korea. Logically, these countries' markets would have a larger influence on the Bitcoin market because they have more weight on itxx. Therefore, we chose to only input these countries' confirmed, recovered, and death patients number into our models to prevent from overfitting our datasets. We obtained these COVID-19 data from [14].

Since there are robots appear on twitter that may post tweets that has "Bitcoin" in their context or hashtags while has nothing to do with the Bitcoin market, we manipulate our data in the following way to reduce noise in our dataset:

Polarity_score_weighted = polarity_score * (retweet amount) * (like amount + 1) * (replies amount)

Tweets post by robots rarely have likes, retweets, or replies. Therefore, by doing the multiplication above, we can give those tweets that are not likely to be post by robots more weights and ensure that we listen to the real voice of the market.

Since the COVID-19 dataset only records data per day, our dataset can only achieve the granularity of days and thus we have limited data points. In the meantime, we have considered many different attributes and also the lag of time. These make overfitting easy to happen. Therefore, we decided to use Principal Component Analysis (PCA) to reduce attributes inputted into models to prevent overfitting. Indeed, in the end, we proved that PCA does improve the performance of our models.

IV. RELATED WORK

A. Prediction objective

We formulate our prediction models in two objectives:

(1) 1-day log return prediction: We predicted the return of a day. This is a regression problem, so we used root mean square error (RMSE) as our evaluation criteria.

(2) price movement prediction: We predicted whether the close price in the next day would be higher than the close price of the day we are looking at. This is a classification problem, so we used accuracy as our evaluation criteria.

B. Models

We used the following machine learning models for prediction: (1) Random Forest; (2) Decision Tree; (3) AdaBoost; (4) Support Vector Machine. All the models are implemented using scikit-learn [15].

C. Input Features

We used the following feature sets as our model inputs: (1) Historical Bitcoin exchange data; (2) Historical Bitcoin exchange data + COVID-19 data (recovery, confirmed, death); (3) Historical Bitcoin exchange data + Twitter data; (4) Historical Bitcoin exchange data + COVID-19 data (recovery, confirmed, death) + Twitter data.

D. Parameters

We consider the lag of time by including different windows of historical data into our models, including (1) Last 5 days data; (2) Last 10 days data; (3) Last 15 days data; We split the dataset into training and testing by having data from 1st January 2020 to 30th June 2020, as the training set and data from 1st July 2020 to 31st July 2020, as the testing set. Five-fold cross-validation is used to tune the hyperparameters of every model. Please see Table 1 for the hyperparameters search space for each model.

Table 1. Hyper-parameter search space for ML models.

Model	Parameter Choices
Random Forest	'max_depth': [5, 10, 15, 20, 25], 'n_estimators': [10, 20, 50, 100, 200]
Decision Tree	'max_depth': [5, 10, 15, 20, 25]
AdaBoost	'max_depth': [5, 10, 15, 20, 25], 'n_estimators': [10, 20, 50, 100, 200]
Support Vector Machine	'C': [0.1, 1, 10], 'gamma': [1, 0.1, 0.01], 'kernel': ['rbf']

E. Results

In this section, we would discuss the result of our models. We ran each model 5 times and took the average of the RMSE scores or classification accuracies because randomness appears while running our models. The RMSE scores for each model are shown in Table 2 and 3 and the accuracies of classification for each model is shown in Table 4 and 5.

Table 2. RMSEs for different models for 1-day log return prediction, with 15 PCA components.

Model	Bitcoin	Bitcoin+COVID	Bitcoin+Twitter	Bitcoin+COVID+Twitter
Lookback time window size: 5 days, PCA = 15 components				
Random Forest	0.0217	0.0217	0.0429	0.0217
Decision Tree	0.0294	0.0371	0.0928	0.0330
AdaBoost	0.0215	0.0221	0.0246	0.0214
Support Vector Machine	0.0226	0.0222	0.0226	0.0222
Lookback time window size: 10 days, PCA = 15 components				
Random Forest	0.0237	0.0235	0.0241	0.0233
Decision Tree	0.0245	0.0266	0.0268	0.0228
AdaBoost	0.0231	0.0258	0.0231	0.0251
Support Vector Machine	0.0226	0.0226	0.0226	0.0226
Lookback time window size: 15 days, PCA = components				
Random Forest	0.0229	0.0229	0.0230	0.0233
Decision Tree	0.0258	0.0295	0.0270	0.0310
AdaBoost	0.0233	0.0231	0.0231	0.0228
Support Vector Machine	0.0226	0.0226	0.0226	0.0226

Table 3. RMSEs for different models for 1-day log return prediction, with 25 PCA components.

Model	Bitcoin	Bitcoin+COVID	Bitcoin+Twitter	Bitcoin+COVID+Twitter
Lookback time window size: 5 days, PCA = 25 components				
Random Forest	0.0220	0.0210	0.0235	0.0222
Decision Tree	0.0215	0.0203	0.0214	0.0256
AdaBoost	0.0222	0.0212	0.0235	0.0221
Support Vector Machine	0.0226	0.0226	0.0226	0.0226
Lookback time window size: 10 days, PCA = 25 components				
Random Forest	0.0222	0.0241	0.0248	0.0242
Decision Tree	0.0216	0.0234	0.0290	0.0229
AdaBoost	0.0229	0.0232	0.0243	0.0233

Support Vector Machine	0.0226	0.0226	0.0226	0.0226
Lookback time window size: 15 days, PCA = 25 components				
Random Forest	0.0235	0.0225	0.0219	0.0224
Decision Tree	0.0237	0.0277	0.0295	0.0255
AdaBoost	0.0225	0.0221	0.0224	0.0229
Support Vector Machine	0.0226	0.0223	0.0226	0.0223

Table 4. Accuracies for different models for price movement prediction, with 15 PCA components.

Model	Bitcoin	Bitcoin+COVID	Bitcoin+Twitter	Bitcoin+COVID+Twitter
Lookback time window size: 5 days, PCA = 15 components				
Random Forest	0.553	0.566	0.546	0.593
Decision Tree	0.600	0.520	0.460	0.510
AdaBoost	0.566	0.573	0.506	0.573
Support Vector Machine	0.566	0.566	0.566	0.566
Lookback time window size: 10 days, PCA = 15 components				
Random Forest	0.586	0.526	0.58	0.513
Decision Tree	0.480	0.473	0.366	0.360
AdaBoost	0.493	0.500	0.526	0.493
Support Vector Machine	0.566	0.566	0.566	0.566
Lookback time window size: 15 days, PCA = 15 components				
Random Forest	0.533	0.513	0.493	0.566
Decision Tree	0.559	0.566	0.546	0.546
AdaBoost	0.466	0.466	0.500	0.440
Support Vector Machine	0.566	0.566	0.433	0.566

Table 5. Accuracies for different models for price movement prediction, with 25 PCA components.

Model	Bitcoin	Bitcoin+COVID	Bitcoin+Twitter	Bitcoin+COVID+Twitter
Lookback time window size: 5 days, PCA = 25 components				
Random Forest	0.586	0.520	0.593	0.526
Decision Tree	0.519	0.493	0.760	0.526
AdaBoost	0.533	0.486	0.740	0.493
Support Vector Machine	0.566	0.533	0.566	0.6
Lookback time window size: 10 days, PCA = 25 components				
Random Forest	0.533	0.539	0.513	0.553
Decision Tree	0.506	0.473	0.373	0.373
AdaBoost	0.473	0.46	0.559	0.506
Support Vector Machine	0.566	0.566	0.566	0.566
Lookback time window size: 15 days, PCA = 25 components				
Random Forest	0.519	0.500	0.519	0.453
Decision Tree	0.473	0.566	0.493	0.559
AdaBoost	0.493	0.566	0.533	0.566
Support Vector Machine	0.566	0.566	0.433	0.566

From the results, we have the following observations and findings:

For classification, it appears that people do investment by considering the recent 5-day information because when we consider a lag of 5 days, we attain better accuracy overall as compared to a lag of 10 days or a lag of 15 days.

Although Support vector machine is the best performing algorithm for predicting the Bitcoin return rate given these four algorithms, it is not very suitable for this type of prediction because it's prediction accuracy is pretty low and not sensitive to the change of data inputs.

The highest two classification accuracies, 74% and 76%, are obtained respectively from inputting 25 PCA components formed by Bitcoin Historical data and Twitter sentiment data with a lag of 5 days into AdaBoost and inputting 25 PCA components formed by Bitcoin Historical data and Twitter sentiment data with a lag of 5 days into Decision Tree. This infers that Twitter data does help in predicting Bitcoin price as previous researches shown and also reassure that people usually refer to information in 5 days for making investment decisions.

After comparing different input feature sets, we do not see a large improvement in accuracy by adding the COVID-19 data. Adding COVID-19 data and Twitter data to Bitcoin Historical data also does not make a large difference from having only Historical data. One possible reason for this is that we have very limited data regarding COVID-19.

V. CONCLUSION

In this paper, we analyze the performance of four machine learning models on predicting Bitcoin return rate and price trend, using four feature sets as inputs. We have the following discoveries: (1) Twitter data improve the performance of models. (2) People consider the information within 5 days when they make decisions on investments. (3) Support vector machine does not perform well in predicting Bitcoin return or price trend. (4) COVID-19 data does not help improve the prediction. However, we have very limited COVID-19 data, so future research with more COVID-19 data may help confirm if this statement is correct.

The greatest shortage of our research is not having sufficient data to train our models. Moreover, we are not able to perform deep learning algorithms because they required a large amount of data to perform appropriately. In the future, when we have more data about COVID-19 and more detailed information like the policies used to control this pandemic, we may have more choices and can predict cryptocurrency prices more accurately.

REFERENCES

[1] Nakamoto S, Bitcoin A. A peer-to-peer electronic cash system[J]. Bitcoin.–URL: https://bitcoin.org/bitcoin.pdf, 2008.

[2] Jiang W. Applications of deep learning in stock market prediction: recent progress[J]. arXiv preprint arXiv:2003.01859, 2020.

[3] Prajapati P. Predictive analysis of Bitcoin price considering social sentiments[J]. arXiv preprint arXiv:2001.10343, 2020.

[4] Wołk K. Advanced social media sentiment analysis for short‐term cryptocurrency price prediction[J]. Expert Systems, 2020, 37(2): e12493.

[5] Wang Y, Chen R. Cryptocurrency price prediction based on multiple market sentiment[C]//Proceedings of the 53rd Hawaii International Conference on System Sciences. 2020.

[6] Guégan D, Renault T. Does investor sentiment on social media provide robust information for Bitcoin returns predictability?[J]. Finance Research Letters, 2020: 101494.

[7] Topcu M, Gulal O S. The impact of COVID-19 on emerging stock markets[J]. Finance Research Letters, 2020: 101691.

[8] Mazur M, Dang M, Vega M. COVID-19 and the march 2020 stock market crash. Evidence from S&P1500[J]. Finance Research Letters, 2020: 101690.

[9] Ciner C. Stock Return Predictability in the time of COVID-19[J]. Finance Research Letters, 2020: 101705.

[10] Mnif E, Jarboui A, Mouakhar K. How the cryptocurrency market has performed during COVID 19? A multifractal analysis[J]. Finance Research Letters, 2020: 101647.

[11] Lahmiri S, Bekiros S. The impact of COVID-19 pandemic upon stability and sequential irregularity of equity and cryptocurrency markets[J]. Chaos, Solitons & Fractals, 2020: 109936.

[12] Poldi F. TWINT-Twitter Intelligence Tool[J]. URL: https://github.com/twintproject/twint (visited on 10/08/2020), 2020.

[13] 7 Countries with the Most Bitcoin Hodlers. URL: https://medium.com/@biditex/7-countries-with-the-most-bitcoin-hodlers-503b205d926f (visited on 10/08/2020), 2020.

[14] CSSEGISandData (2020) Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE. GitHub repository, Available at: https://github.com/CSSEGISandData/COVID-19 (Accessed: 8 October 2020).

[15] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python[J]. the Journal of machine Learning research, 2011, 12: 2825-2830.