Ethereum Price Prediction Regression Models via Social Media Posts and Bitcoin Price Actions

Beile Y'aaqob I. Aisin
Science Academy
University of Maryland
12/13/2021

Abstract

Social media has long been used as a source of data to make predictions of social events. It has demonstrated exceptional usage in building predictive models for the financial market. This project attempts to combine the natural language data obtained from Reddit and the price actions of Bitcoin, and to test the feasibility of constructing reliable price prediction models. Two machine learning models, Decision Trees and Gradient-Boosted Trees, were trained using the same set of data. The models yield similar performance; however, the results indicate the variety of the data must be significantly improved, in order for the models to generate reliable performances.

Introduction

Social media has long been used by data scientists as a data source to predict social event outcomes. In the past, it had provided data for election result predictions (Tumasjan et al., 2010), movie sale predictions (Asur and Huberman, 2010), public health event predictions (Volkova et al., 2017), and financial market predictions (Martin, 2013).

The effects of social media on the financial market had caught the attention of the mainstream society during the first and second quarters of 2021 -- the AMC Entertainment and Game Stop "short-squeeze" organized by the social media group the Wall Street Bet received major media coverage, and the tweets from Elon Musk regarding Dogecoin have cause massive price actions on the famous meme cryptocurrency. Therefore, this project attempted to use social media posts regarding the blockchain ecosystem Ethereum (ETH), collected from Reddit, as the main dataset to construct machine learning models capable of predicting the price actions of Ethereum's native token, Ether.

Raw Datasets

The main dataset that was used in the project is the *Reddit Ethereum Dataset* obtained from Kaggle (link). This dataset contains two CSV files -- the first file is a collection of Reddit posts relating to Ethereum, which were posted from 07/03/2011 to 11/01/2021, and the second file is a collection of Reddit comments relating to Ethereum, which were posted from 05/25/2021. The two CSV files also have different lengths -- the Reddit post file contains 479,260 rows and the Reddit comment file contains 1,132,331 rows. The columns in the two files are different as well -- for the Reddit post file, the columns are ['type', 'id', 'subreddit.id', 'subreddit.name', 'subreddit.nsfw', 'created_utc', 'permalink', 'domain', 'url', 'selftext', 'title', 'score'], and for the Reddit comments file, the columns are ['type', 'id', 'subreddit.id', 'subreddit.name', 'subreddit.nsfw', 'created_utc', 'permalink', 'body', 'sentiment', 'score'].

The price actions of several other cryptocurrencies were also included as supporting datasets. The price actions of Bitcoin (BTC) were included as training and testing data. Although BTC and ETH have fundamental differences (BTC was intended for financial transactions and store-of-value, while ETH was intended to power the blockchain projects built on the Ethereum ecosystem), they are the two oldests cryptocurrencies as well as the two biggest by market capitalization. The price actions of the two cryptocurrencies, Cardano (ADA) and Solana (SOL), which are directly competing with ETH were included as supporting data too. However, both ADA and SOL are relatively newer cryptocurrency projects, which means their oldest timestamps are not as old as *the Reddit Ethereum Dataset*'s. Due to this reason, the price actions of the two were not included in the training-testing process. Nevertheless, their most recent price actions were visualized alongside the price actions of ETH and BTC for direct comparisons.

Data Preparation

The raw datasets were processed before forming a single dataset for the test-train split for machine learning training. The price actions data are daily, while the Reddit dataset may contain no entry or numerous entries for each day. To solve this problem, the Reddit dataset was converted to daily as well. The method is to combine all the rows with texts posted on the same day. Then the Reddit dataset was merged with the BTC price actions dataset by dates. Also, it needs to be noted that only the CSV file containing the Reddit comments were used in the project, because PySpark kept reading many columns in the Reddit posts CSV file as 'null', despite heavy efforts to fix the issue. Due to the time limitation of the project, the Reddit post CSV file had to be abandoned, in order to meet the deadline.

Because the objective of the machine learning models was to predict ETH prices, a random test-train split was not feasible -- the order of the data must follow the timeseries. Therefore, the training data were manual selected, as early as 2016-05-15 and as late as 2019-10-07. Within this time frame, the cryptocurrency market completed a full bull and bear market cycle (**Fig. 1**). The testing data was from 2019-10-08 to 2021-07-12; in this time frame, both ETH and BTC completed another parabolic price increase and a significant price decrease. The data from 2021-07-13 to 2021-11-01 (the last date of the dataset) were ingested after the training-testing process to record the performances of the models with wild data - data that are not included in the test-train process.

Algorithms & Machine Learning Architecture

There were two machine learning models trained in total -- both are regression models, in order to forecast future prices. The first model was trained using a Decision Trees regressor, and the second model was trained using a Gradient-Boosted Trees (GBT) regressor. Before the training data were ingested for training, a 4-stage Spark Pipeline was utilized for data transformation. Stage-1 is to tokenize the input text (Reddit posts), and Stage-2 is filtering the stopwords in the output of Stage-1. After filtering the stopwords, Stage-3 converts the data from text format into a numerical vectors with vector size equals to 100 via the Word2Vec function in Spark, and then the output of Stage-3 was assembled with the BTC price action data into a single matrix, as the last step before being ingested into the regression models for training.

Evaluations

Unlike classification models, regression model evaluations generally do not measure overall accuracy of model performances; instead, three other methods were used to evaluate the models - Root Mean Square Error (RMSE), Mean Square Error (MSE), and Coefficient of Determination (R^2).

	Decision Trees	**GBT**
RMSE	918.71	924.47
MSE	545.65	544.41
R^2	-0.01	-0.03

Table 1: performances of both models

The two models yielded similar performances. However, all of the evaluations indicate that the models are far from ideal. Much more work needs to be done in order to predict the price actions of ETH.

Analyses and Recommended Future Work

There are several potential explanations for the poor performances of the models, and they are mostly related to the datasets.

First, the quality of the dataset itself may not be good. As the word cloud shows (**Fig. 2**), most of the vocabularies appearing in the Reddit dataset have no differences with the English used in daily languages. Only a selected few of them are related to blockchain and cryptocurrency. A possible solution to this problem is to increase the size of the stopword list, which means to filter out much more vocabularies and to leave out more related words. Also, the Reddit data does not have data available for every day, for example, the next entry after 2016-05-23 is 2016-05-26.

Second, the data lacked variety -- there were only two input columns for the test-train process. In reality, to reach optimal performance, the dataset is recommended to have much more features.

Related to the second explanation, the field of blockchain technology and cryptocurrency is still in its infancy. In other words, it is not yet evolved enough to produce sufficient data to build regression models with high performances. In the original design of this project, the price actions of many ETH competitors were planned to be included in the test-train process. However, most of them are insufficient in time frame to match with the Reddit data -- their price actions were absent from the 2016 to 2019 market cycle. Moreover, because the industry is still young, the ETH competitors have not yet developed they action patterns different than ETH, and ETH itself, although having fundamental differences with BTC, still follows the general price trends of BTC (**Fig.1, Fig 3a, 3b, 3c**, & **3d**).

Conclusion

Although social media data have long been used to make predictions about social events, we do not yet have sufficient data to build reliable models for cryptocurrency price predictions. Many social events mentioned in the introduction have existed in human society for decades, if not for centuries. The industry of blockchain and cryptocurrency, on the other hand, is barely ten years old. Much more time is needed for it to evolve.

Bibliography

Asur, S., and Huberman, B. A.. (2010). Predicting the future with social media. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 01, 492– 499. IEEE Computer Society

Glenski, M., Weninger, T., & Volkova, S.. (2019). Improved Forecasting of Cryptocurrency Price using Social Signals. *Not Published*.

La Morgia, M., Mei, A., Sassi, F., & Stefa, J.. (2021). The Doge of Wall Street: Analysis and Detection of Pump and Dump Cryptocurrency Manipulations. *Submitted (not yet published)*.

Martin, V. (2013). Predicting the French Stock Market Using Social Media Analysis. *2013 8th International Workshop on Semantic and Social Media Adaptation and Personalization*.

Mirtaheri, M., Abu-El-Haija, S., Morstatter, F., Steeg, G. V., & Galstyan, A.. (2019). Identifying and Analyzing Cryptocurrency Manipulations in Social Media.

Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe, I. M.. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, volume 10, 178–185. AAAI.

Volkova, S., Ayton, E., Porterfield, K., & Corley, C. D.. (2017). Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *Plos One, 12*(12).

Appendix:

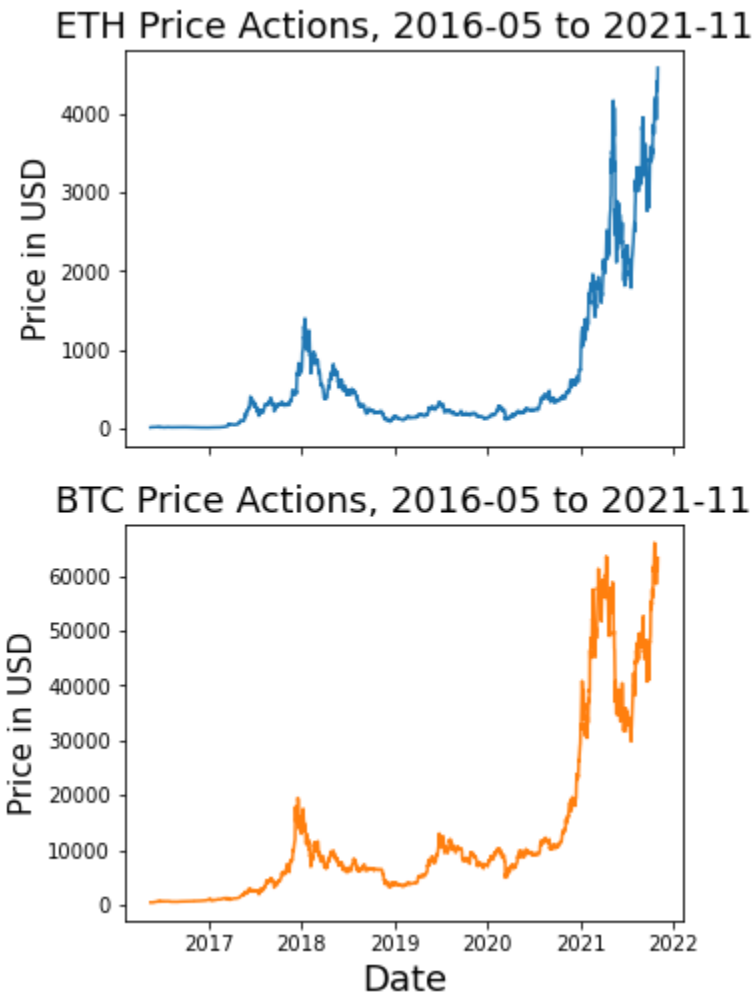

Fig.1: price actions of ETH and BTC from 2016 to 2021

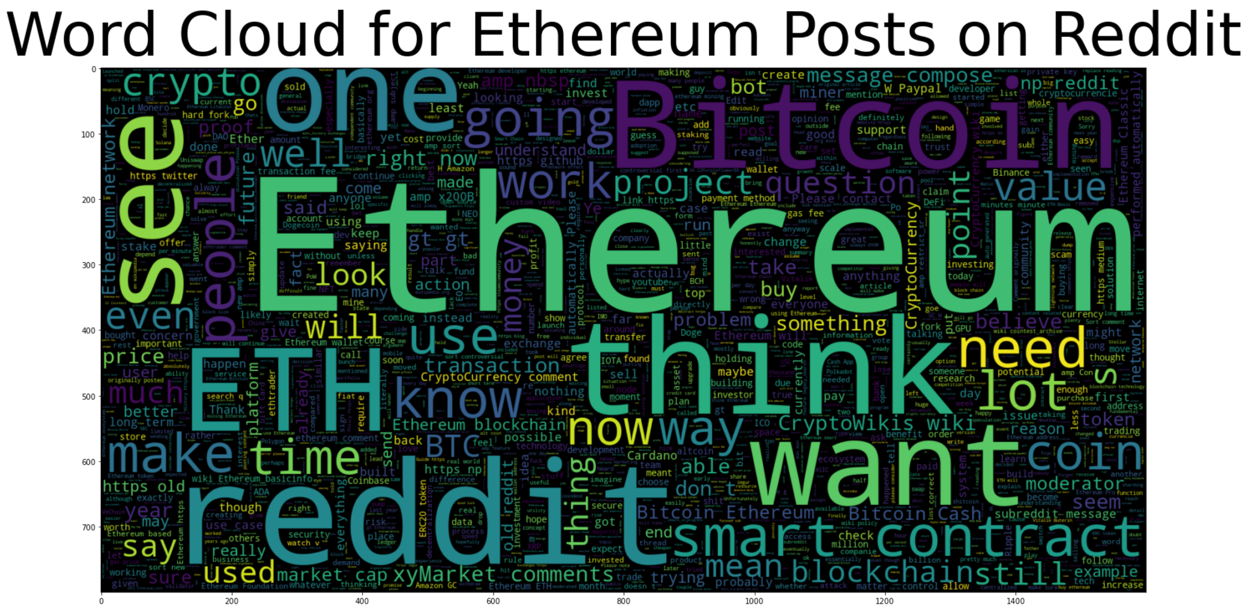

Fig. 2: most common vocabularies appeared in the Reddit dataset

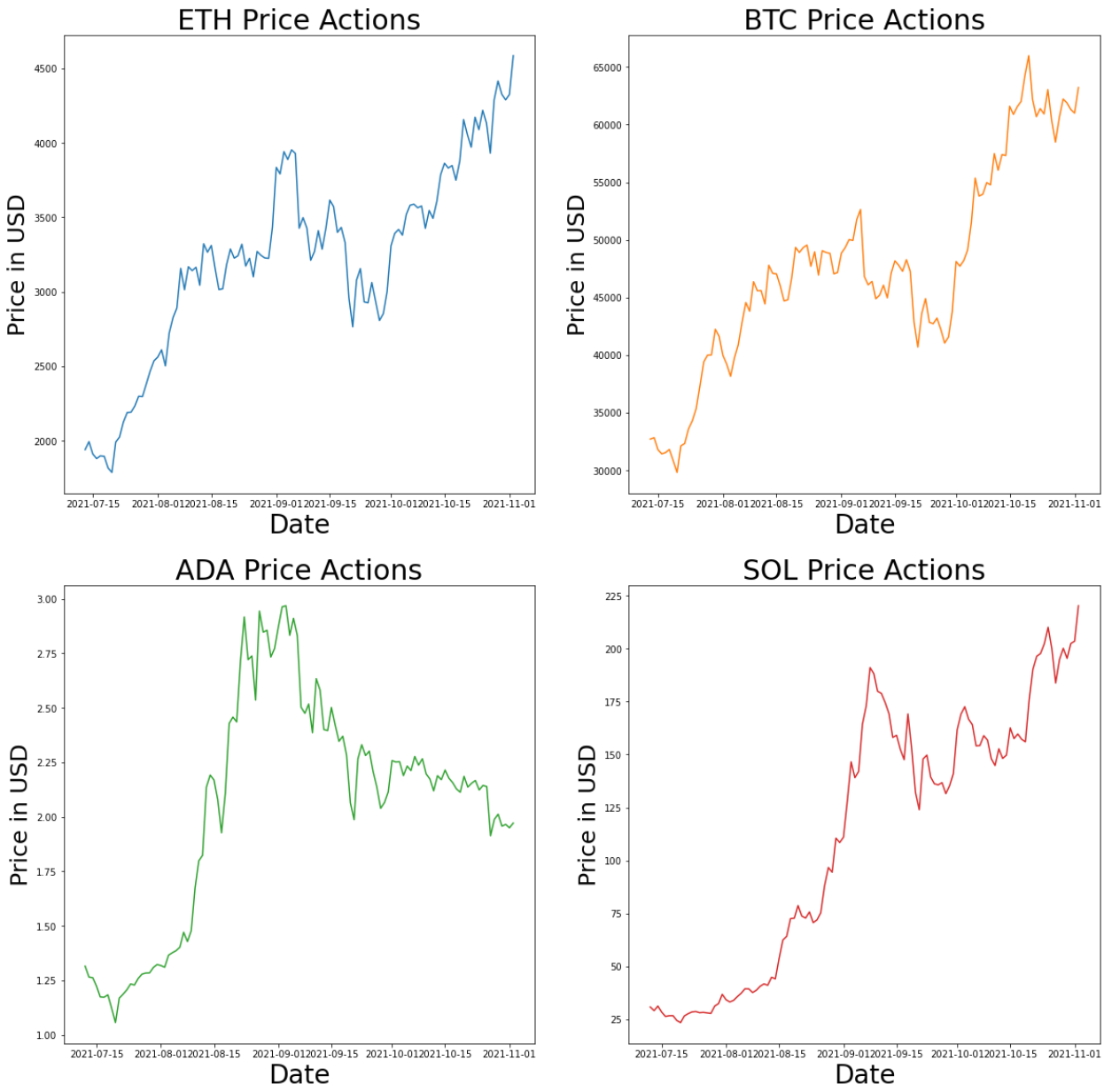

Fig. 3: the price actions of **a)** ETH, **b)** BTC, **c)** ADA, and **d)** SOL, in the recent months.
(from left to right and up to down)