Use of Historical and Social Media Data to Predict Bitcoin Price

# INTRODUCTION

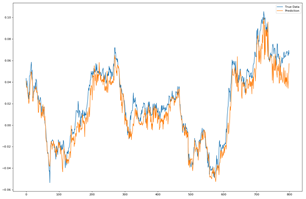
With the near-daily news reports of record-high bitcoin prices, and the pervasiveness of social media, and the seemingly direct relationship between the two topics, this connection seemed like a good topic to investigate via neural networks. The goal of this project was to determine if it is possible to predict the direction of price change (not the price itself) of Bitcoin, given certain historical trading data and a subset of social media data from the same time period. Bitcoin price would appear to be particularly susceptible to social media sentiment given that it is unregulated, has no ‘PR Room’ to release news reports, has no press releases, etc. Social media seems to be the main channel that information regarding the Bitcoin market is spread., implying that social media can help predict the trend in Bitcoin price.

.

1. PRIOR WORK

Implementations of neural nets to predict stock prices are easy to find. However, results are obviously sub-par. If it was that easy to predict a stock price, with precision, ‘everybody would be doing it.’ A casual examination of studies of this type (prediction of stock price in the future) quickly reveal a common pattern. That is, the predictive model predicts that tomorrow’s stock price will be the actual price of the stock today. Authors then claim a low error rate and explain how successful the model is.

Predictions of Stock Prices [1]



Many ‘Predictive Models’ seem to follow this same pattern. Rather than following this same pattern, I decided to investigate the possibility of predicting the direction of price change, rather than the price itself, with relation to Bitcoin.

# DATA

Two avenues of data collection were used for this project. First, historical data regarding Bitcoin trades was collected. This data closely resembles standard stock-market data. Data collected in this category includes the daily opening price of Bitcoins, daily closing price, daily high, daily low, volume traded, and market capitalization of Bitcoin. Data of this type was available, free of charge, from April 28, 2013 until December 12, 2017. Older data on price and other aspects of bitcoin trading was also available, but this was the largest set of free, complete, data that I could find. This date range gives 1690 data points. With regards to social media, the twitter tweet history was searched for the hashtag #bitcoin. For each date in the date range given by the limited set of historical data collected, 100 tweets containing the hashtag #bitcoin were returned. To attempt to locate influential tweets, search results were limited to tweets designated by twitter as ‘Top Tweets’ were requested. In certain situations, fewer than 100 tweets were returned for a given date. This phenomenon was more common the farther back in time the target date was.

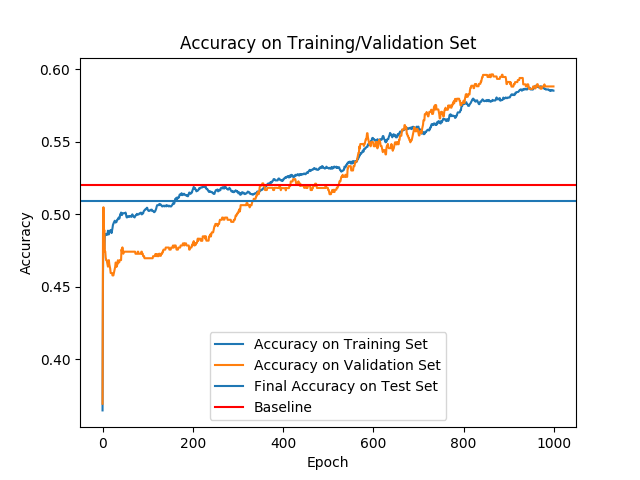
After gathering the tweets via twitter search via a modified version of the code found at [2], the body of the tweet was examined to determine the sentiment of the tweet. Sentiment Analysis was done via the Vader module in SciKit Learn. Vader claims to match the state of the art in sentiment analysis accuracy. It combines tone (positive/negative) with intensity (really/just a bit) to generate distinct positive, negative, and combined scores for each tweet. For this project, I decided to use the combined rating (positive score – negative score) as the category of tweet. The neural net was given the above historical data, plus two more columns, one contain the number of ‘positive’ tweets, and another containing the number of negative tweets that I had scraped for that date. Altogether, input to the neural network consisted of 1690 inputs, with each input consisting of eight separate pieces of information, opening price, closing price, daily high, daily low, volume, market cap, number of positive-sentiment tweets, and number of negative-sentiment tweets.

# Experiment Parameters

The neural net for this experiment uses LSTM cells, in an attempt to capture trends that may exist in the data. Two hidden LSTM layers were used, each of size 64. These layers fed directly to the output layer. The tensorflow adamoptimizer was used to anneal control learning rates. Mini-batches of size 5 were used, in combination with a time-step value of 5 for a total of 25 data points used for each mini-batch. A sliding, non-overlapping time-window was used to select data points. Cross-fold validation was used to reduce overfitting. Data was split in a 70-20-10, with 20 percent used for validation and 10 percent used for testing. The

# Results

The network was trained for 1,000 epochs. Based on the results of the validation set, this seemed like a good place to stop training. Accuracy on the validation set reached just below .60. The baseline for the data was 0.520118. This accuracy could be achieved by simply guessing that the price will increase every day. Final results are reflected in the following chart.



While the training and validation set accuracy appeared to indicate good progress, the average final accuracy on the previously-unseen data folds was actually below the baseline. Accuracy on some of the test splits were comparable to the validation accuracy, however, this was highly variable. Some splits showed an accuracy of greater than 0.6, while accuracy on other splits was between 0.2 and 0.3. However, averaged together to get a representative value of model accuracy, the results were lower than expected.

# DISCUSSION

Possible sources of error in this study include inaccurate data and a non-representative subset of tweets. Historical data for bitcoin transactions are freely available online. However, data from different sources sometimes did not agree. Since bitcoin is non-centralized, this might be caused by variations among the different bitcoin 'exchanges'.

Also of concern is the limited number of tweets collected. Since access to the twitter 'firehose' was unavailable, number of available tweets was limited. Time required to scrape 100 tweets per day for each day in the historical price data was acceptable, but did not provide much useful data. Upon closer inspection, many of the tweets gathered did not appear to provide any useful sentiment, whatsoever. A common tweet throughout the data gathered was what appeared to be bots tweeting the current bitcoin price. To actually get useful information from tweets, a much greater volume of tweets probably must be collected.

References

*:*

1. http://www.jakob-aungiers.com/articles/a/Multidimensional-LSTM-Networks-to-Predict-Bitcoin-PriceUSA: Abbrev. of Publisher, year, ch. *x*, sec. *x*, pp. *xxx–xxx.:*
2. https://github.com/Jefferson-Henrique/GetOldTweets-pythonW.-K. Chen, *Linear Networks and Systems.* Belmont, CA: Wadsworth, 1993, pp. 123–135.