Bitcoin Stock Prediction Using Deep Learning and Sentiment Analysis

Balaji Mudaliyar | Neelam Babel
mudaliyar.b@husky.neu.edu
babel.n@husky.neu.edu
CSYE 7245 Big Data Sys & Intelligence Analytics, Spring 2018 Northeastern University

Abstract

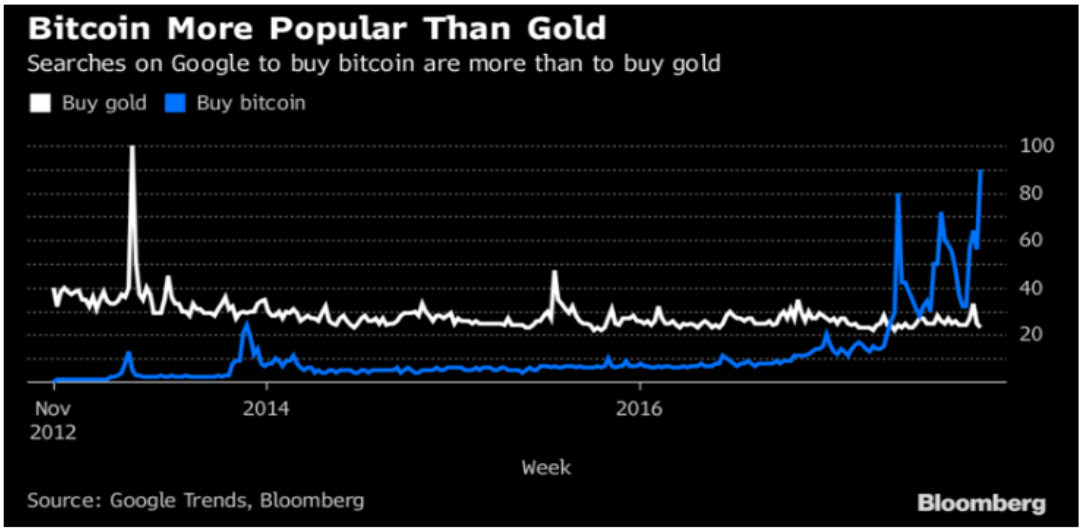
Crypto currencies have gained a lot of public attention over the last few years. They are primarily characterized by fluctuations in their price and number of transactions. Such unstable fluctuations have served as an opportunity for speculation for some users while hindering most others from using or investing in crypto-currencies. Hence, as part of this project, our main aim was to predict the bit-coin closing prices as accurately as possible based on the given historical data. Since it is a time series data, we are using Recurrent Neural Network using LSTM. Furthermore, we extended our approach to analyze bit-coin related tweets for sentiment fluctuations to understand it's impact on price change in the near future and combined it with the original model to see if we can improve upon the accuracy. Based on our observations, RNN (LSTM) generates a reliable model with a loss of ~0.00014583 just after 100 epochs and the sentiment scores were helpful indicators as well. However, analysis needs to be extended to more number of tweets so that the model can be trained to avoid over-fitting.

Introduction

It's 2018, the people of the world generate 2.5 million terabytes of information a day. 500 million tweets, 1.8 billion pieces of shared information on Facebook, each and every day. Twitter specifically has become known as a location where news is quickly disseminated in a concise format.

When regarding a financial commodity, the public confidence in a particular commodity is a core base of its value. Social media has served as platform to express opinions since their inception, and as such tapping into the open APIs provided of the likes of Facebook and Twitter, these arguably biased pieces of information become available with a sea of meta-data.

Bit-coin (BTC), the decentralized cryptographic currency, is similar to most commonly known currencies in the sense that it is affected by socially constructed opinions; whether those opinions have basis in facts, or not. Since the Bitcoin was revealed to the world, in 2009, it quickly gained interest as an alternative to regular currencies. As such, like most things, opinions and information about Bitcoin are prevalent throughout the Social Media sphere.



The motive behind this research is that there are a number of design flaws in Bitcoin, and people are trying to invent new coins to overcome these defects hoping their inventions will eventually replace Bitcoin. To June 2017, the total market capital of all cryptocurrencies is 102 billion in USD, 41 of which is of Bitcoin. Therefore, regardless of its design faults, Bitcoin is still the dominant cryptocurrency in markets. As a result, many altcoins cannot be bought with fiat currencies, but only be traded against Bitcoin. Hence, we chose Bitcoin as our commodity.

Dataset

The datasets is used from https://www.quandl.com/

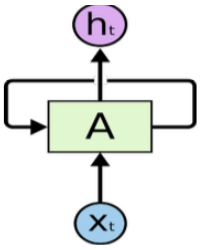
Quandl.com-
Dataset include the Adjusted Open, Adjusted High, Adjusted Low, Adjusted Close, Adjusted Volume for BTC, Adjusted Volume for Currency and Weighted Price for Bitcoin retrieved using Quandl's free Bitcoin API for dates ranging from January 7, 2014 to April 21, 2018.

Model Architecture Design

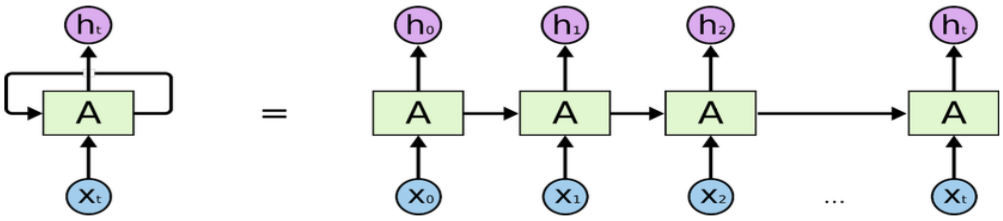
Our first framework is a **Recurrent Neural Network using LSTM** (Long Short Term Memory) trained on 3 popular stock market indicators and past prices as key data points to find an optimal technique for cryptocurrency stock market prediction.
Our second framework includes **Sentimental Analysis** score calculated from twitter comments coupled with past prices as key data points to make the predictions for future closing prices.
Comparison will be made based on their performance. Both techniques have some advantages and disadvantages. Our research will analyze advantages and limitations of these techniques to find which technique is comparatively better specifically for Bitcoin stock market prediction.
Also, prediction will be done using Random forest algorithm and it will be a forecast for the next 1 week.

Recurrent Neural Networks

Recurrent neural networks are networks with loops in them, allowing information to persist.



A recurrent neural network can be thought of as multiple copies of the same network, each passing a message to a successor.

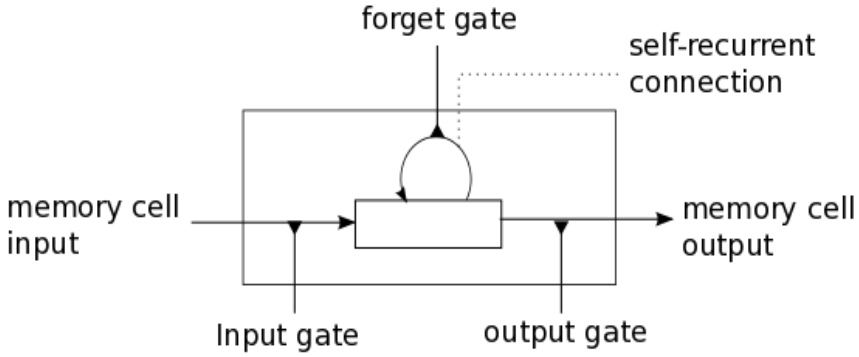


RNN's might be able to connect with previous information to the present task, i.e. using previous video frames might inform the understanding of the present frame.

Problems with RNN

RNN is trained by back propagation through time, and therefore unfolded into feed forward net with multiple layers. When gradient is passed back through many time steps, it tends to grow or vanish, commonly popularly known as **Vanishing/Exploding Gradient**.

To overcome this, we use **LSTM** which introduces a new structure called a memory cell. A memory cell is composed of four main elements: an input gate, a neuron with a self-recurrent connection (a connection to itself), a forget gate and an output gate. The self-recurrent connection has a weight of 1.0 and ensures that, barring any outside interference, the state of a memory cell can remain constant from one time step to another.



Sentiment Analysis Approach to Cryptocurrency Speculation

We have analyzed user sentiments related to crypto currencies on social media, e.g., Twitter, and converted the feedback into scores. We then study the correlation between the scores and fluctuations in price and trade volume to determine any relation.
Articles on crypto currencies, such as Bit coin, are rife with speculation these days, with hundreds of self-proclaimed experts advocating for the trends that they expect to emerge. So, we felt that analysis of the current trending tweets for the term Bit-coin to predict its closing price for the next day seemed like the most unbiased approach to resolving the biased opinions strewn around the web.
This approach also resonates with our personal approach to track the closing prices of crypto currencies we have invested in. We always find ourselves skimming through the trending tweets on the matter.

Sentimental Analysis
Sentiment Analysis is the process of 'computationally' determining whether a piece of writing is positive, negative or neutral. It's also known as **opinion mining**, deriving the opinion or attitude of a speaker.

Why sentiment analysis?
Business: In marketing field companies use it to develop their strategies, to understand customers' feelings towards products or brand, how people respond to their campaigns or product launches and why consumers don't buy some products.
Politics: In political field, it is used to keep track of political view, to detect consistency and inconsistency between statements and actions at the government level. It can be used to predict election results as well!
Public Actions: Sentiment analysis also is used to monitor and analyse social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.

Three major steps are done in program:
-Authorize twitter API client.
-Make a GET request to Twitter API to fetch tweets for a particular query.

```
# We create an extractor object:
extractor = twitter_setup()

# We create a tweet list as follows:
tweets = extractor.user_timeline(screen_name="@Bitcoin", count=200)
print("Number of tweets extracted: {}.\n".format(len(tweets)))

# We print the most recent 5 tweets:
print("5 recent tweets:\n")
for tweet in tweets[:5]:
    print(tweet.text)
    print()
```

```
Number of tweets extracted: 200.
```

-Parse the tweets. Classify each tweet as positive, negative or neutral and calculate score which is SA score ranging from -1 to 1.

```
from textblob import TextBlob
import re

def clean_tweet(tweet):
    '''
    Utility function to clean the text in a tweet by removing
    links and special characters using regex.
    '''
    return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)", " ", tweet).split())

def analize_sentiment(tweet):
    '''
    Utility function to classify the polarity of a tweet
    using textblob.
    '''
    analysis = TextBlob(clean_tweet(tweet))
    if analysis.sentiment.polarity > 0:
        return analysis.sentiment.polarity
    elif analysis.sentiment.polarity == 0:
        return 0
    else:
        return analysis.sentiment.polarity
```

```
# We construct lists with classified tweets:

pos_tweets = [ tweet for index, tweet in enumerate(data['Tweets']) if data1['SA'][index] > 0]
neu_tweets = [ tweet for index, tweet in enumerate(data['Tweets']) if data1['SA'][index] == 0]
neg_tweets = [ tweet for index, tweet in enumerate(data['Tweets']) if data1['SA'][index] < 0]
```

```
# printing percentages:

print("Percentage of positive tweets: {}%".format(len(pos_tweets)*100/len(data1['Tweets'])))
print("Percentage of neutral tweets: {}%".format(len(neu_tweets)*100/len(data1['Tweets'])))
print("Percentage de negative tweets: {}%".format(len(neg_tweets)*100/len(data1['Tweets'])))
```

```
Percentage of positive tweets: 37.0%
Percentage of neutral tweets: 54.0%
Percentage de negative tweets: 9.0%
```

Random Forest

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it's simplicity and the fact that it can be used for both classification and regression tasks.

Random Forest is a supervised learning algorithm. The forest it builds, is an ensemble of Decision Trees, most of the time trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.

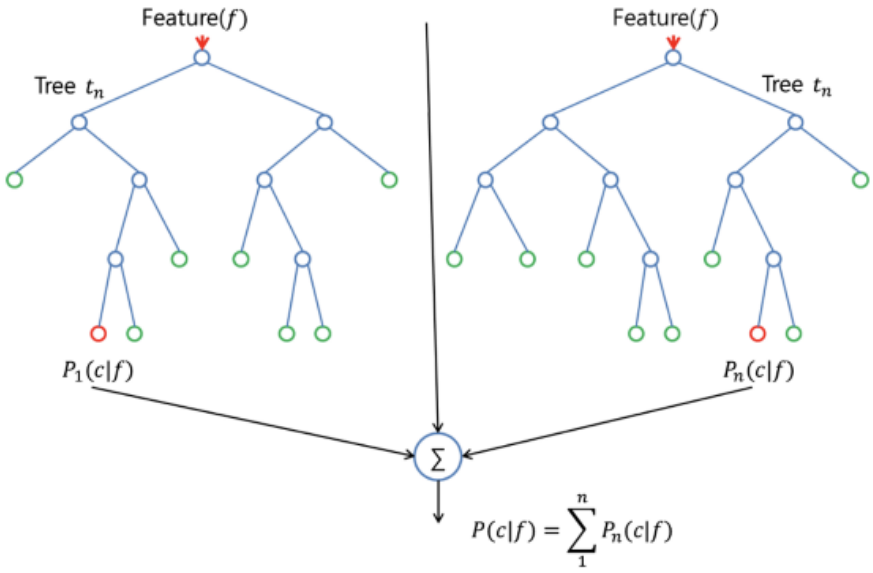
One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems.

Important Hyperparameters:

1. I**ncreasing the Predictive Power:**

   n_estimators hyperparameter, which is just the number of trees the algorithm builds before taking the maximum voting or taking averages of predictions. In

general, a higher number of trees increases the performance and makes the predictions more stable, but it also slows down the computation.



2. **Increasing the Models Speed:**
   random_state makes the model's output replicable. The model will always produce the same results when it has a definite value of random_state and if it has been given the same parameters and the same training data.

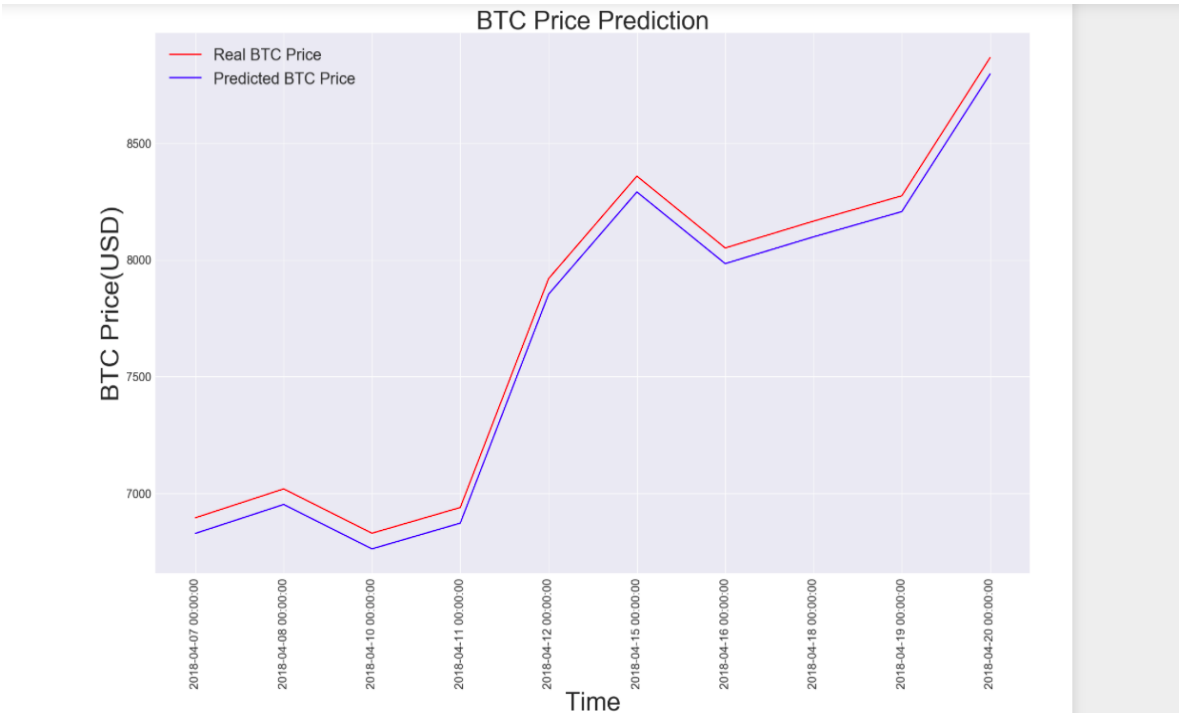
Code with Documentation

GitHub Link: The complete code with documentation can be found on the below link

https://github.com/balaji-mudaliyar/Bitcoin-Predicition/blob/master/Bitcoin%20closing%20price%20prediction%20using%20Deep%20Learning%2C%20Random%20Forest%20and%20Sentiment%20Analysis_Project.ipynb

Results

Based on our observations, we can infer that Recurrent Neural Network using LSTM when trained on Initial Features (Historical Bitcoin Prices) and Secondary Features (Fiat Currency Stock Market Technical Indicators**)** generates a reliable model with a loss of ~0.0001458 after 100 epochs and does a satisfactory job in predicting the BTC prices. Below are the results when we try to predict the BTC prices for the period – 07 April'2018 – 20 April 2018 after training the model on historical BTC data of previous 4 years.



Furthermore, on extending our approach to Sentiment Analysis to couple sentiment scores with prices, we were able to add variance to our data and that resulted in a better prediction of the future BTC price as we had expected.
We have used Random Forest classifier to predict the bit-coin prices by collecting tweets over one week. We are feeding the data for the period – 18 April'208 to 21 April'2018 to the model to predict the closing price for 22 April'2018.

Date	SA	Average Open Price (USD)	Average High Price (USD)	Average Low Price (USD)	Average Close Price (USD)
2018-04-18	0.16	7891	8254	7872	8165
2018-04-19	-0.08	8165	8299	8092	8274
2018-04-20	0.01	8272	8936	8220	8865
2018-04-21	0.03	8866	9043	8616	8811
2018-04-22	0.07	0	0	0	0

The prediction using random Forest and sentimental analysis for 22nd April was **8821.03 USD** and the actual closing price was **8802.46 USD**.
Actual closing price of 22nd Apri 2018 from coinmarketcap:

Apr 23, 2018	8,794.39	8,958.55	8,788.81	8,930.88	6,925,190,000	149,448,000,000
Apr 22, 2018	8,925.06	9,001.64	8,779.61	8,802.46	6,629,900,000	151,651,000,000
Apr 21, 2018	8,848.79	8,997.57	8,652.15	8,895.58	7,548,550,000	150,337,000,000

Acknowledgment
We would like to show our gratitude to professor **Nik Bear Brown** for guiding us during this project.

Discussion

Despite their numerous strengths, neural networks suffer from a number of weaknesses that researchers should keep in mind.
Sample size: Neural networks need larger samples in order to be estimated properly. This is due to the large number of parameters introduced in such models that link the inputs to the hidden neurons, which are then linked to the output variable. As the data available for the training and testing of the model increase, one would then expect the marginal gains in accuracy to increase. There is no rule of thumb rule for the "optimal" sample size for which one can expect neural nets to improve noticeably.
Tweet Count Limit: For Sentiment Analysis, we have used Tweepy API to extract the data from twitter, limiting it to a count of 200 for testing our approach. Increasing the count should lead to more specific results.

References

[1] Reid F, Harrigan M. An analysis of anonymity in the bitcoin system: Springer; 2013.
[2] Böhme R, Christin N, Edelman B, Moore T. Bitcoin: Economics, technology, and governance. The Journal of Economic Perspectives. 2015;29(2):213–38.
[3] Nakamoto S. Bitcoin: A peer-to-peer electronic cash system. 2008.
[4] Kondor D, Pósfai M, Csabai I, Vattay G. Do the rich get richer? An empirical analysis of the Bitcoin transaction network. PloS one. 2014;9(2):e86197 doi: 10.1371/journal.pone.0086197 [PMC free article] [PubMed]
[5] Ron D, Shamir A. Quantitative analysis of the full bitcoin transaction graph Financial Cryptography and Data Security: Springer; 2013. p. 6–24.
[6] Garcia D, Tessone CJ, Mavrodiev P, Perony N. The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy. Journal of the Royal Society Interface. 2014;11(99):20140623. [PMC free article] [PubMed]
[7] Kondor D, Csabai I, Szüle J, Pósfai M, Vattay G. Inferring the interplay between network structure and market effects in Bitcoin. New Journal of Physics. 2014;16(12):125003.
[8] Kristoufek L. BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. Scientific reports. 2013;3. [PMC free article] [PubMed]
[9] Kristoufek L. What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysis. PloS one. 2015;10(4):e0123923 doi: 10.1371/journal.pone.0123923 [PMC free article] [PubMed]
[10] Yelowitz A, Wilson M. Characteristics of Bitcoin users: an analysis of Google search data. Applied Economics Letters. 2015;22(13):1030–6.