Cryptocurrency Price Speculation & Prediction

Project 2: Milestone 4 Final

Thip Rattanavilay
Bellevue University, DSC-680

Cryptocurrency Price Speculation & Prediction

Abstract

I have spent the entire year of 2020 understanding Blockchain and Cryptocurrency, especially Bitcoin and Ethereum. These two crypto are the most volatile markets today and has gained a lot of attention from investors across the globe. Cryptocurrency, being a novel technique for transaction system, has led to a lot of confusion among the investors and any rumors or news on social media has been claimed to significantly affect the prices of cryptocurrencies. The goal of this study is to predict prices for crypto coins using Machine Learning Techniques for the few I weeks for project 2 I will prepare a strategy to maximize gains for investors, friends and of course myself. I also aim to find out if there is a co-relation between fluctuating crypto prices and related news.

Cryptocurrency Price Speculation & Prediction

Introductions

Stock market is one of the most volatile data available in terms of Machine learning datasets. Researchers have been long trying to predict the stock market and any breakthrough in this field would result in, literally, the people being able to mint money. Cryptocurrencies, to be specific, has gained a lot of traction in the recent years from investors across the globe. Cryptocurrency being a novel technique for transaction system has led to a lot of confusion among the investors and any rumors or news on social media has been claimed to significantly affect the prices of cryptocurrencies. Bitcoin is one of the oldest and biggest cryptocurrencies being traded as of now, in terms of the volume being traded. It is so big that even now, with the advent of thousands of new cryptocurrencies, Bitcoin has a market share of more than 55% as compared to other cryptocurrencies, being followed by Ethereum at 8.57%. This says a lot about why Bitcoin might be a really interesting and important stock to predict. Also, Bitcoin prices fluctuate heavily. Over the past 2 years, Bitcoin has seen its highest price around $20000 and its lowest price around $900. It is very sporadic, and this is one of the most important reasons which attracted us to analyze and predict its price. Our rest of the paper would discuss our various attempts to predict its price and us trying to reason out how external factors like news affect its price.

Business Problem

Cryptocurrency, especially Bitcoin, is one of the most volatile markets today and has gained a lot of attention from investors across the globe. Cryptocurrency, being a novel technique for transaction system, has led to a lot of confusion among the investors and any rumors or news on social media has been claimed to significantly affect the prices of cryptocurrencies. The goal of this study is to predict prices for Bitcoin using Machine Learning Techniques for the next day and prepare a strategy to maximize gains for investors. We also aim to find out if there is a co-relation between fluctuating Bitcoin prices and related news.

Dataset

For the purpose of this research, I will utilize data from Jan-2018 to Aug-2019, concerning the hourly prices in USD and were divided into training set consisting of data from Jan-2018 to Feb-2019 (10176 values) and testing set from Mar-2019 to Aug-2019 (4416 values). This data was taken from Kaggle.com, kraken.com website and others like Coinbase etc., which is a trading platform and data repository for cryptocurrency exchanges. Also, I will utilize those data for forecasting and predictions (number of past prices taken into consideration).

Variables:
Open

Cryptocurrency Price Speculation & Prediction

Close
Volume
Volume Currency
Weighted Price Date

Crypto Variables:
BTC
ETH
LTC
XRP
ETC
STR
DASH

We use Coinbase dataset from the publicly available Coinbase site which is by far the most reliable crypto exchange in United States. Our horizon forecast is days and therefore we aggregate the hourly data by day.

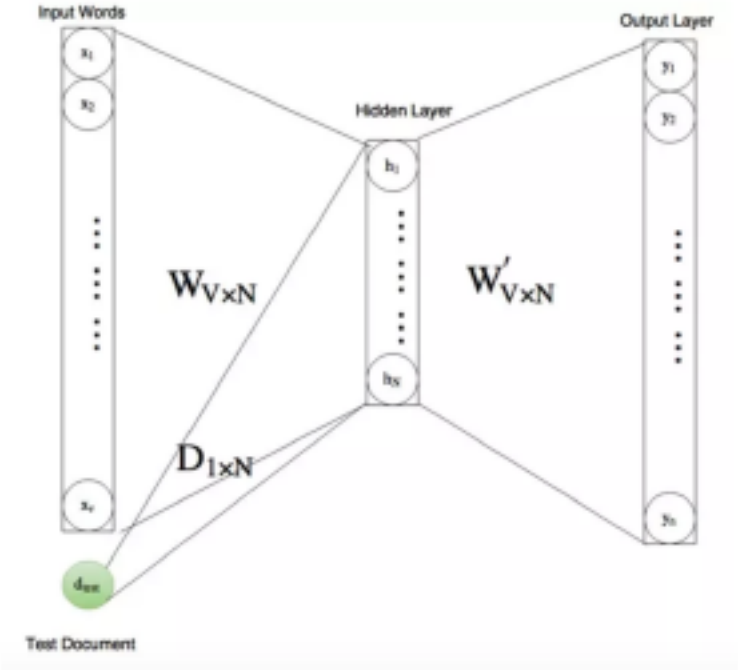
NAN Values: Common technique in time series data to handle NAN values is to forward fill the detain case of missing values. We need to be sure to not use the backward fill strategy in this case as we don't want to predict the past given the future.

Before we can jump into modeling, it's best to get an idea of the structure and ranges by making a few exploratory plots. This will also allow us to look for outliers or missing values that need to be corrected. Some of the interesting trends.

<u>Tweets and news data</u>

Now, to investigate the claim of bitcoin related news and tweets affecting the bitcoin prices, we needed to perform some natural language processing on the input data. We started off with the input texts and converted them to word embeddings. Word embeddings are basically vectorial representation for words. We used the Doc2Vec utility (based off of Word2Vec) from the genism library for generating these word vectors.

We then feed these sentence vectors to the models. To get the previous x day's news, we use the same kind of shifting as with the bitcoin price data to get the previous x day's news. A lot of days in our dataset had no news and this too was an important negative signal for the model, which it picked up on. Figure 1



(Figure 1: A representation of the internal Doc2Vec neural net)

Methods

To solve the problem using Machine learning, we first tried to categorize the problem and tried to find previous solutions on how they solved it. We quickly learned that, since the problem involves prices which are changing with time, this could be modelled as a Series prediction problem. In parallel, we also tried to solve the problem as a normal machine learning problem with the features being the previous prices and the output being the price predicted for that day. We explain what models we have used and how we configure them to predict the Bitcoin prices.

ARIMA

Autoregressive integrated moving average (ARIMA) is a statistical regression model, which can be utilized in time series forecasting applications, such as finance. ARIMA makes predictions while considering the lagged values of a time series, while accommodating for non-stationarity. The model, which is one of the most popular linear models for time series forecasting, originates from the autoregressive (AR) and moving average (MA) models, as well as their combination, also known as ARMA. For making predictions with the ARIMA model, we had to follow a step-by-step procedure to be able to feed the data to the ARIMA model. This first involved Visualizing the time series data: It is essential to analyze the trends prior to building

any kind of time series model. The details we are interested in pertains to any kind of trend, seasonality, or random behavior in the series.

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time
and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

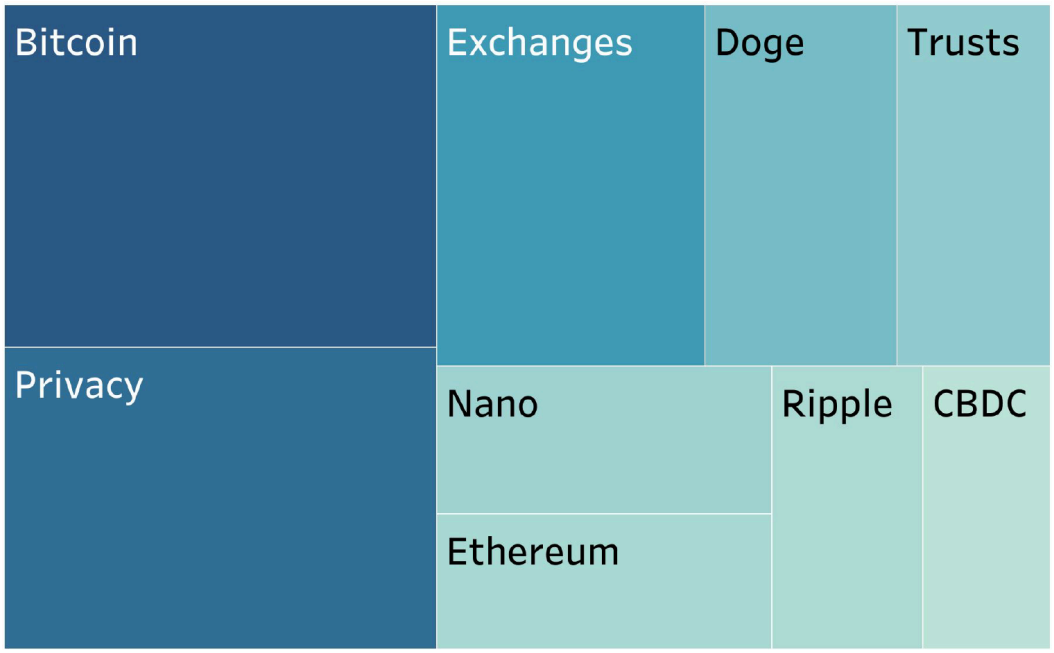
This bootstrapping procedure leads to better model performance because it decreases the variance of the model, without increasing the bias. This means that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated. Simply training many trees on a single training set would give strongly correlated trees (or even the same tree many times, if the training algorithm is deterministic); bootstrap sampling is a way of de-correlating the trees by showing them different training sets.

Recurrent Neural Net

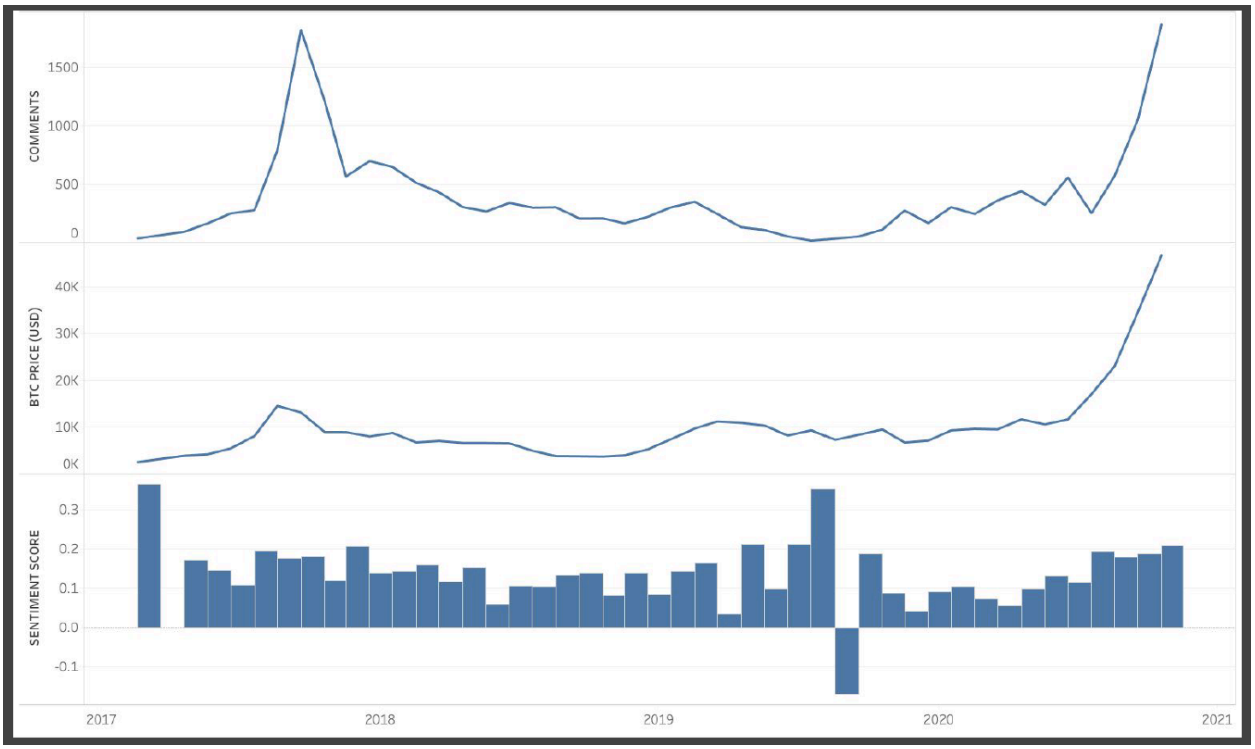
RNNs provide a generalization of the feed forward network model for dealing with sequential data, with the addition of an ongoing internal state ht serving as memory buffer for processing sequences. In this paper, we employ a seq2seq RNN. More concretely, let fx1; ….; xng define the set of n sequence inputs of a seq2seq RNNs and fy1; ….; yng be the set of outputs. For this study the internal state of the network is processed by Gated Recurrent Units (GRU).

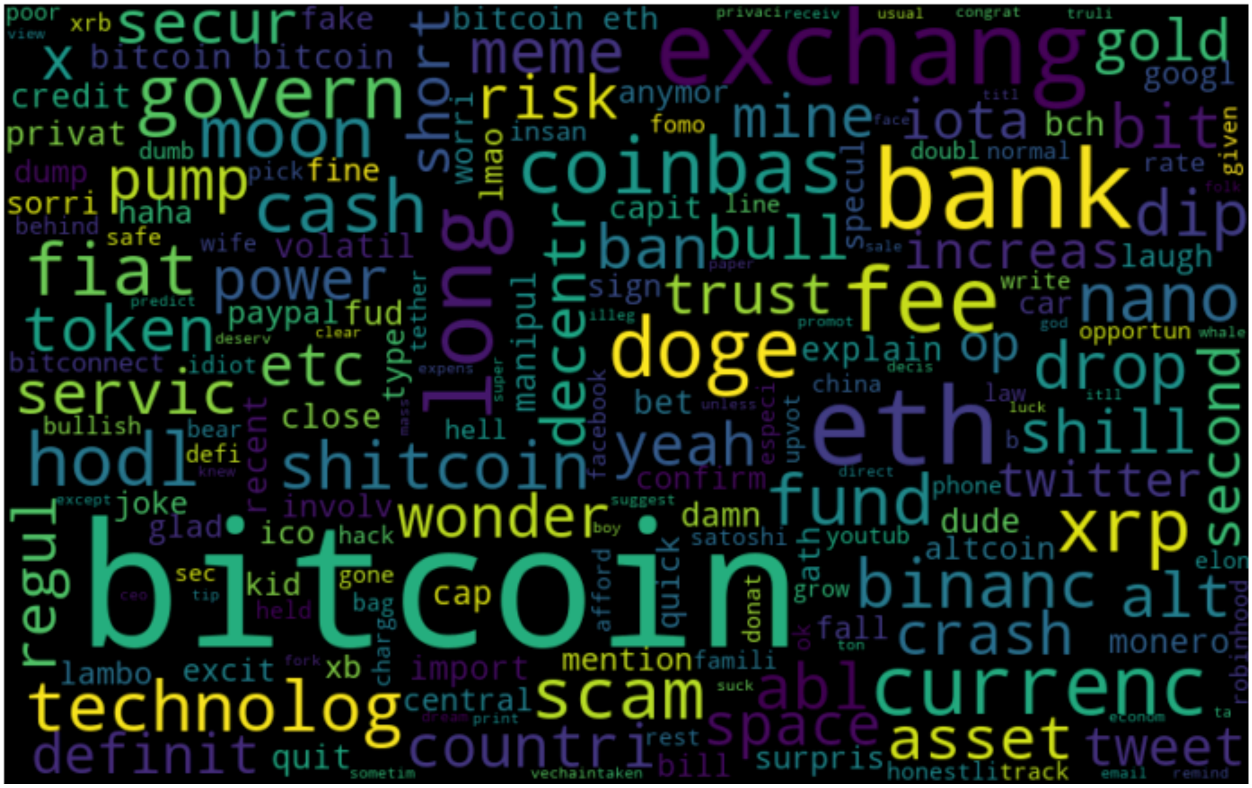
Illustrations

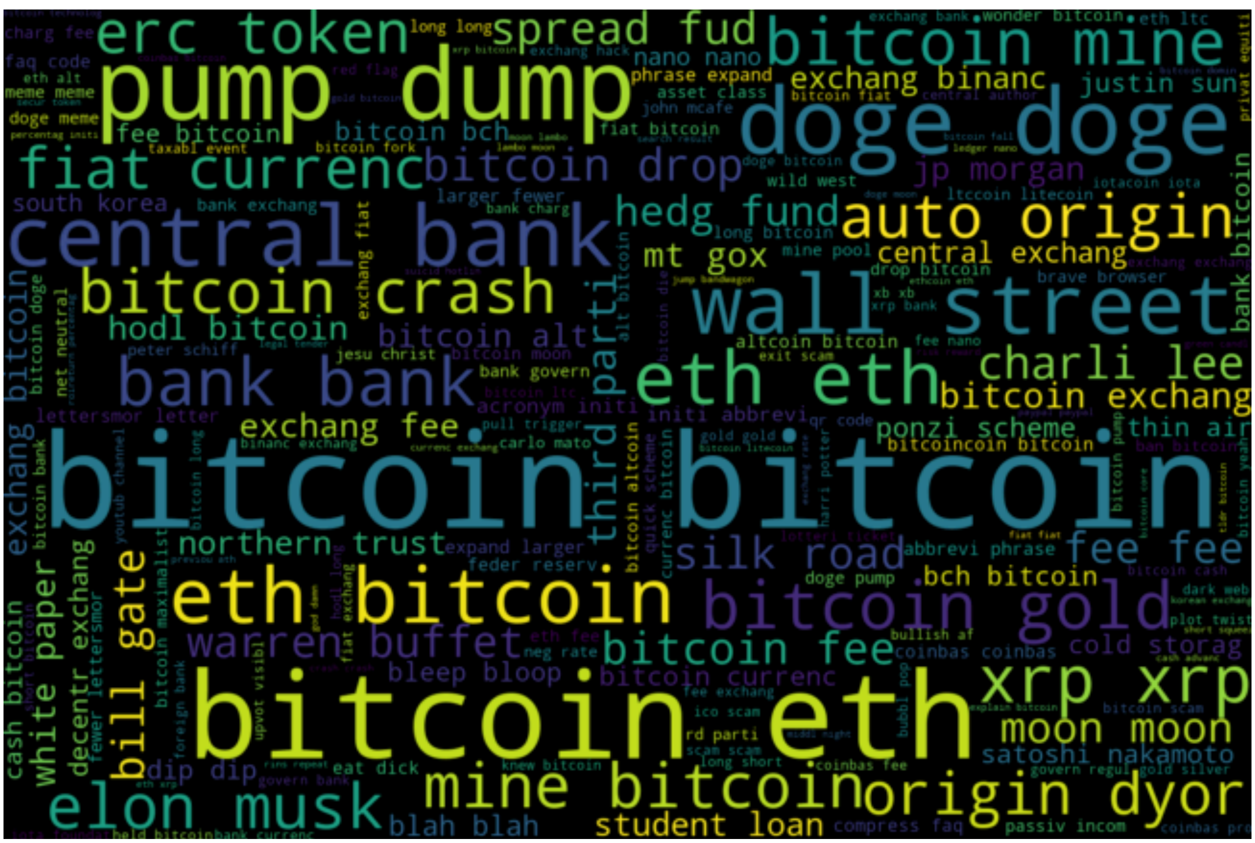
Top crypto coins on the market



Sentiment Twitter posts

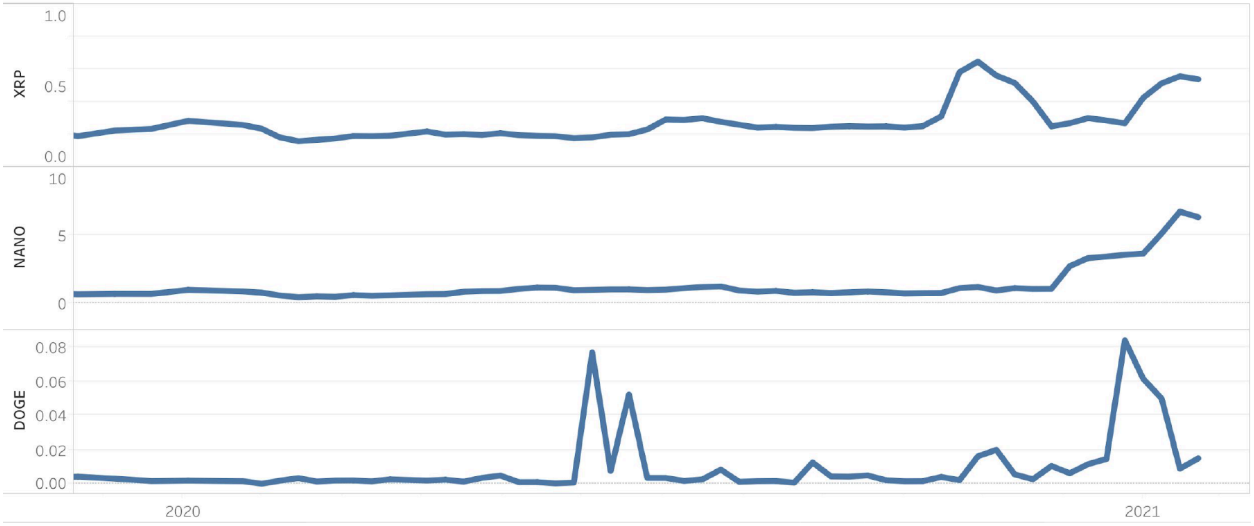



bitcoin
exchang
bank
eth
long
fee
doge
coinbas
govern
secur
moon
cash
fiat
token
hodl
shitcoin
technolog
currenc
scam
asset
xrp
fund
binanc
crash
shill
drop
nano
trust
bull
ban
decentr
short
meme
risk
mine
iota
gold
wonder
regul
servic
etc
pump
power
yeah
twitter
second
alt
countri
definit
space
tweet

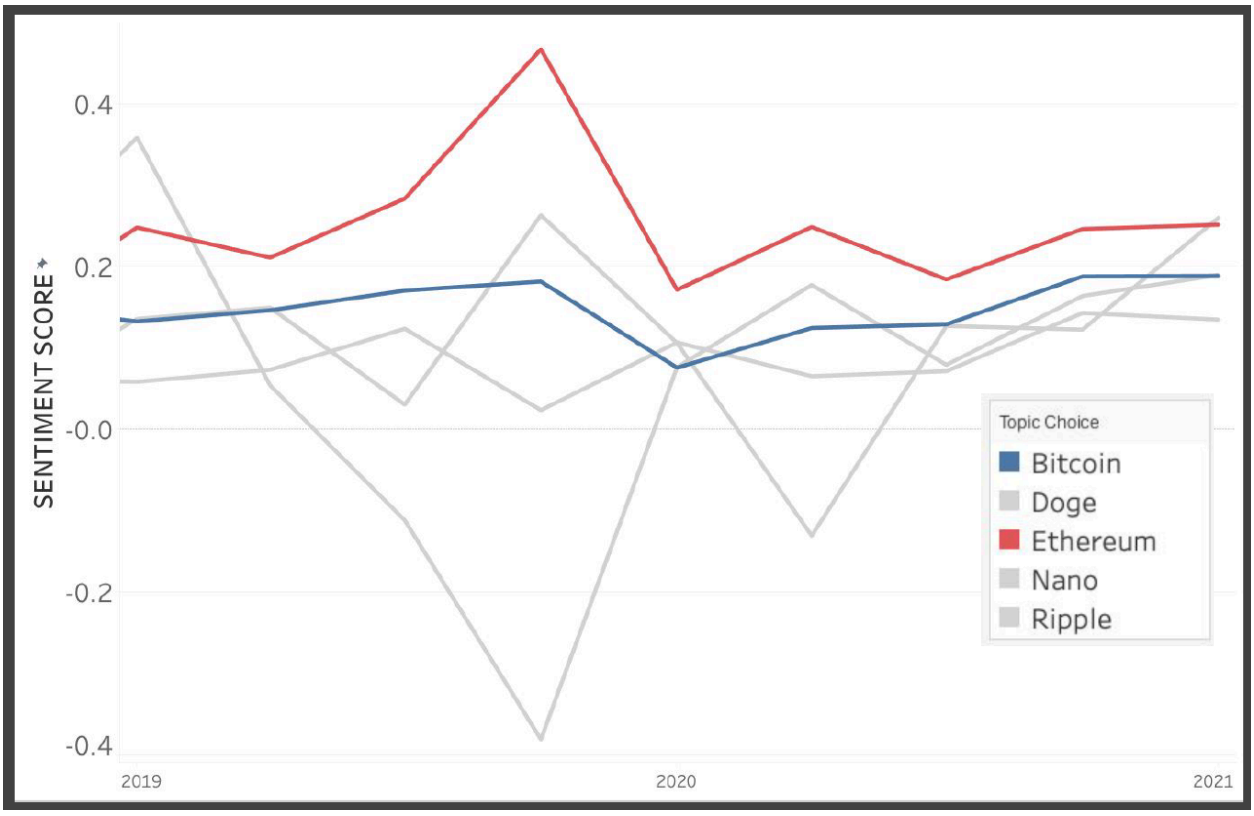

bitcoin bitcoin
bitcoin eth
pump dump
doge doge
central bank
wall street
erc token
bitcoin mine
spread fud
fiat currenc
bitcoin drop
auto origin
bitcoin crash
hedg fund
eth eth
bank bank
eth bitcoin
bitcoin gold
elon musk
mine bitcoin
origin dyor
xrp xrp
silk road
northern trust
warren buffet
bitcoin fee
charli lee
bitcoin exchang
hodl bitcoin
bitcoin alt
exchang fee
third parti
bill gate
white paper
moon moon
satoshi nakamoto
jp morgan
mt gox
central exchang
exchang binanc
justin sun
ponzi scheme
thin air
fee fee
bch bitcoin
cold storag
student loan
blah blah
bleep bloop

Cryptocurrency Price Speculation & Prediction

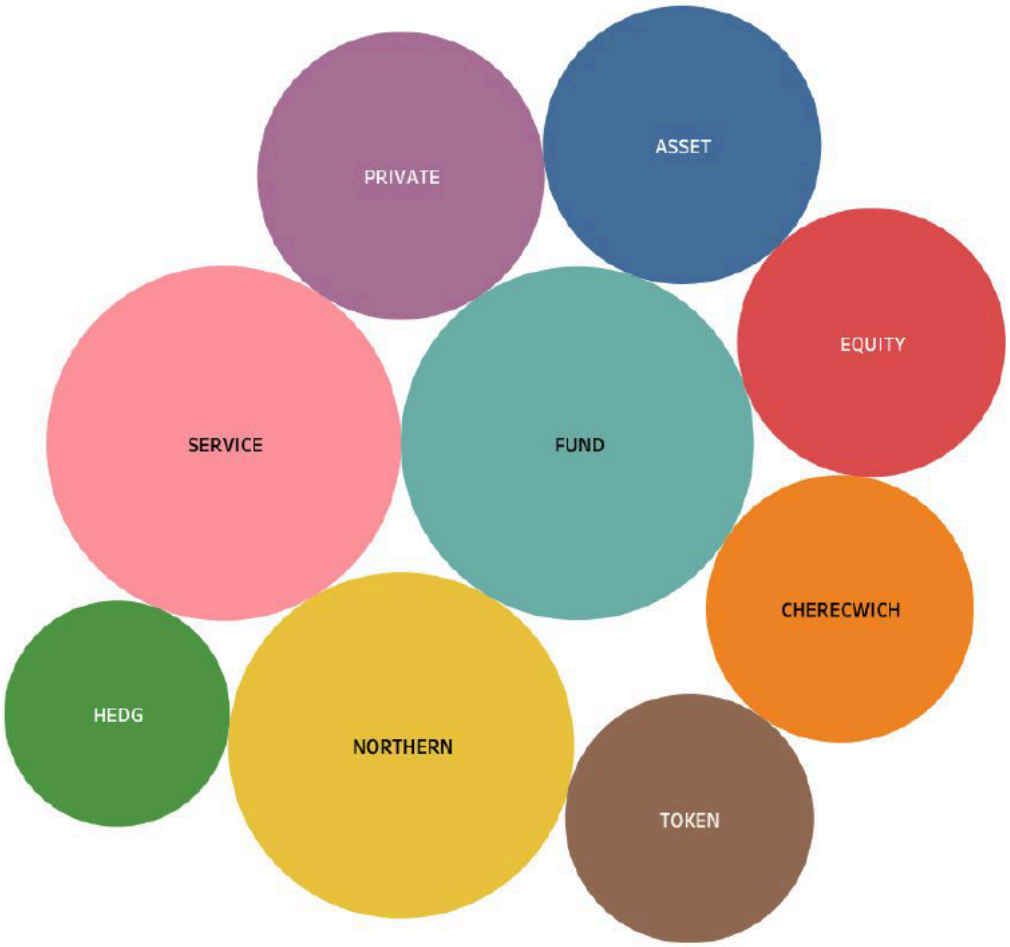
Risky Investments



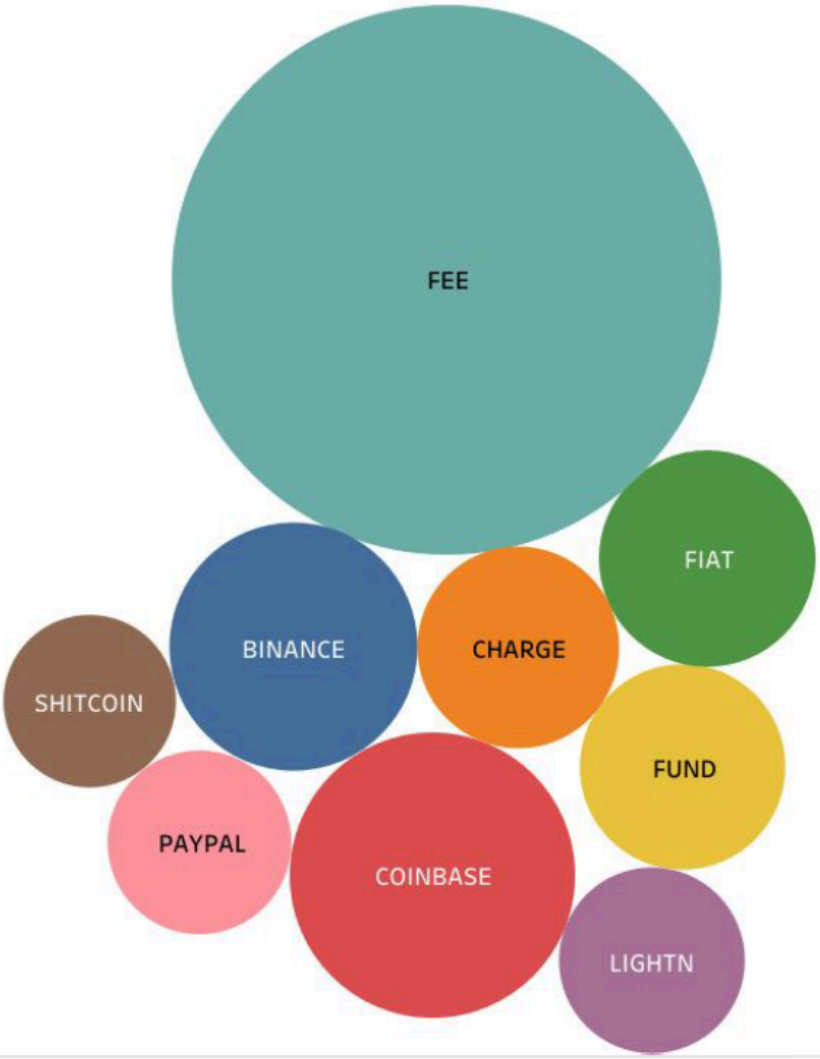
Bitcoin vs Ethereum vs Doge vs Nano vs Ripple



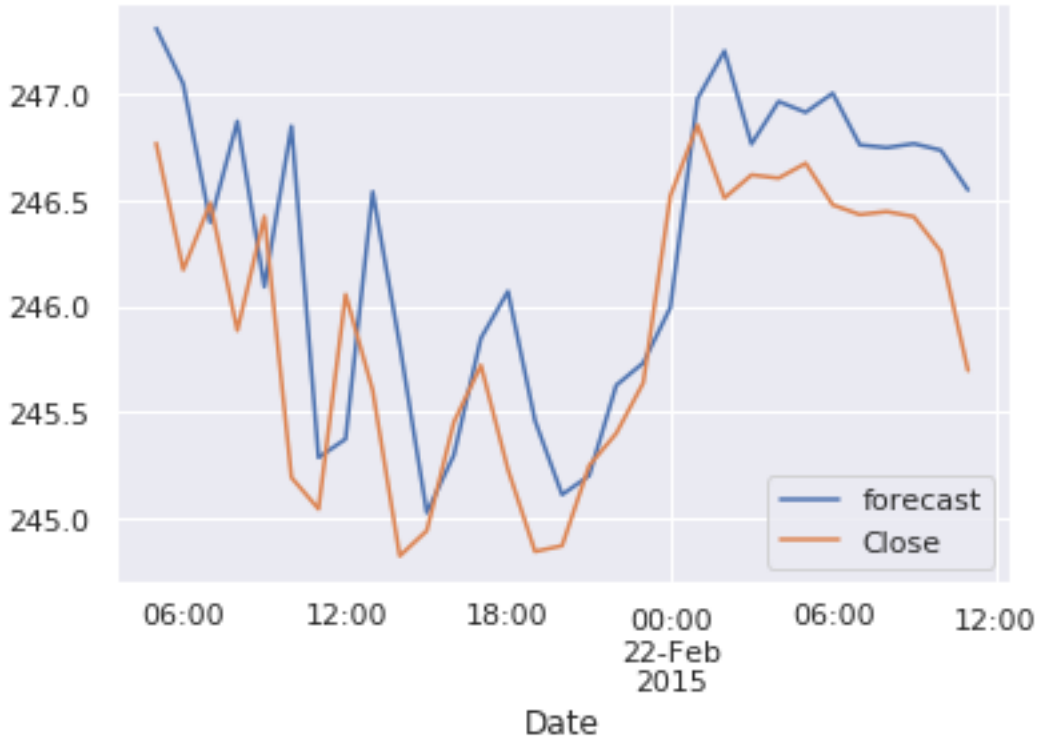
Cryptocurrency Price Speculation & Prediction

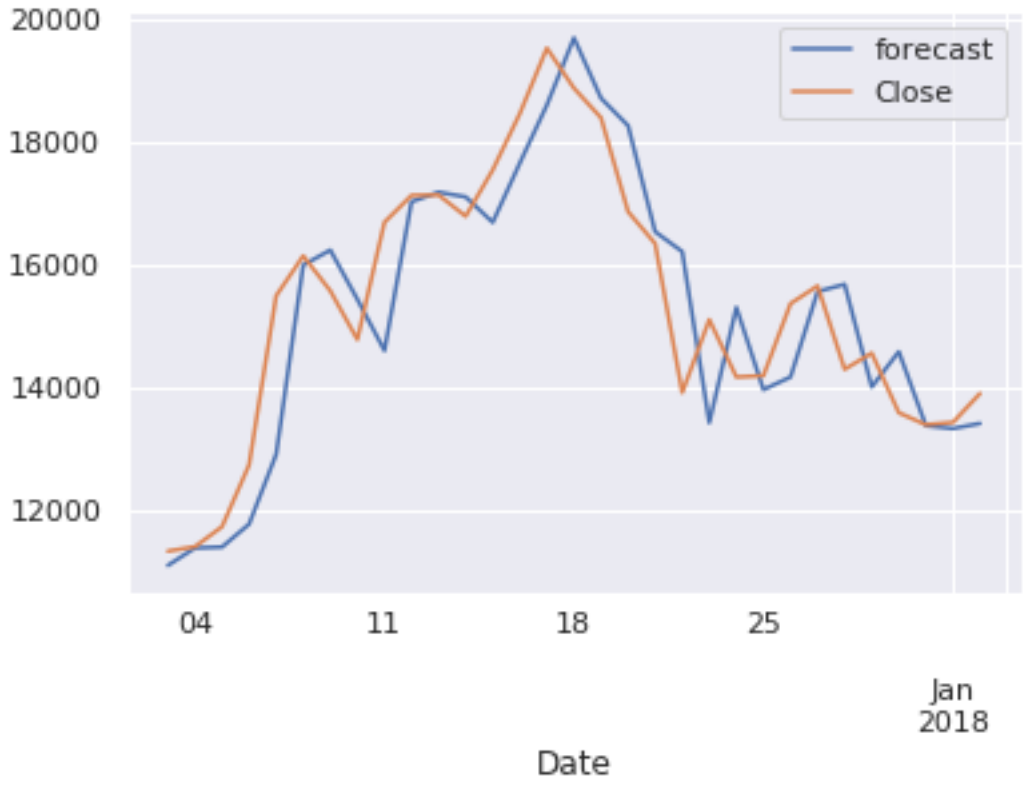


Cryptocurrency Price Speculation & Prediction




forecast
Close
20000
17500
15000
12500
10000
7500
5000
2500
0
2015
2016
2017
2018
2019
Date


247.0
246.5
246.0
245.5
245.0
forecast
Close
06:00
12:00
18:00
00:00
22-Feb
2015
06:00
12:00
Date



ARIMA performed similar to other models when the window length was 3 and forecasted a single day's value (present). This was expected given how ARIMA works which is intuitively autoregression for the past 3 values. However, the model failed significantly when it had to predict forecasts for the next 7 days.

Random Forest performed in a pretty standard way when the window length was 3 and forecasted a single day's value (present). The RMSE with Random Forest was decent and we saw that increasing the max depth up to 16 gave us the best results. Increasing the max depth up to 50 gave us the best results when considering the news data. We saw that Random forests started to over fit when we decreased the number of trees, and this was expected.

RNN The best architecture of RNN we found to be predict rmse to be decent in all cases was a deep RNN with 7 LSTM. We used mean squared error as loss function and for optimizer, we used Adam optimizer and learning rate of 0.01 as opposed to 0.007 of default value. We also tried training using GRU and got very similar rmse results. However, with GRU, the training time reduced with little accuracy loss. We found training with 1000 epochs was sufficient enough. Adding to that, we found the predictions to be most accurate for immediate next day and relatively lesser with each day but better than ARIMA model. (Figure 2)

Model	RMSE
ARIMA	$395.24
RF	$406.23
RF with News	$849.61
RNN	$271.62
RNN with News	$539.85

Figure 2: Compassions of different models

Appendix

The input consists of a list of past Bitcoin data with step size of 256. The output is the predicted value of the future data with step size of 16. Note that since the data is ticked every five minutes, the input data spans over the past 1280 minutes, while the output cover the future 80 minutes. The datas are scaled with Min MaxScaler provided by sklearn over the entire dataset. The loss is defined as Mean Square Error (MSE).

Model	#Layers	Activation	Validation Loss	Test Loss (Scale Inverted)
CNN	2	ReLU	0.00029	114308
CNN	2	Leaky ReLU	0.00029	115525
CNN	3	ReLU	0.00029	201718
CNN	3	Leaky ReLU	0.00028	108700
CNN	4	ReLU	0.00030	117947
CNN	4	Leaky ReLU	0.03217	12356304
LSTM	1	tanh + ReLU	0.00007	26649
LSTM	1	tanh + Leaky ReLU	0.00004	15364
GRU	1	tanh + ReLU	0.00004	17667
GRU	1	tanh + Leaky ReLU	0.00004	15474
Baseline (Lag)	-	-	-	19122
Linear Regression	-	-	-	19789

1. Python
2. Tensorflow
3. Keras
4. Pandas
5. Numpy
6. h5py
7. sklearn

BTC_Close	BTC_Volume	ETH_Close	ETH_Volume
434.33	36278900.0	0.948024	206062
433.44	30096600.0	0.937124	255504
430.01	39633800.0	0.971905	407632
433.09	38477500.0	0.954480	346245
431.96	34522600.0	0.950176	219833

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_1 (LSTM)                (None, 7, 512)            1058816
_________________________________________________________________
dropout_1 (Dropout)          (None, 7, 512)            0
_________________________________________________________________
lstm_2 (LSTM)                (None, 7, 512)            2099200
_________________________________________________________________
dropout_2 (Dropout)          (None, 7, 512)            0
_________________________________________________________________
lstm_3 (LSTM)                (None, 512)               2099200
_________________________________________________________________
dropout_3 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 513
_________________________________________________________________
activation_1 (Activation)    (None, 1)                 0
=================================================================
Total params: 5,257,729
Trainable params: 5,257,729
Non-trainable params: 0
```

Cryptocurrency Price Speculation & Prediction

10 Questions

What are the causes dips and spikes in crypto values?

The causes are due to news media, Tweets and Reddit comments.

How should I invest?

Based on the finding it is best to invest into a respectable cryptocurrency like Bitcoin or Ether and leave meme coins alone.

If I invest into crypto what are the risk?

The risk is that cryptocurrency is hard to predict the outcome and that it's not stable. But if you properly invest you can be a millionaire overnight.

What is cryptocurrency?

A cryptocurrency, crypto-currency, or crypto is a collection of binary data which is designed to work as a medium of exchange wherein individual coin ownership records are stored in a ledger which is a computerized database using strong cryptography to secure transaction records, to control the creation of additional coins, and to verify the transfer of coin ownership.

What is the purpose of cryptocurrency?

The purpose is to have a decentralized cryptocurrency and not govern by the big bank in a centralized control environment.

What is blockchain?

A blockchain is a growing list of records, called blocks, that are linked together using cryptography. Each block contains a cryptographic hash of the previous block, a timestamp, and transaction data (generally represented as a Merkle tree). The timestamp proves that the transaction data existed when the block was published in order to get into its hash. As blocks each contain information about the block previous to it, they form a chain, with each additional block reinforcing the ones before it. Therefore, blockchains are resistant to modification of their data because once recorded, the data in any given block cannot be altered retroactively without altering all subsequent blocks.

Is cryptocurrency taxable?

Cryptocurrency is considered "property" for federal income tax purposes, meaning the IRS treats it as a capital asset. This means the crypto taxes you pay are the same as the taxes you might owe when realizing a gain or loss on the sale or exchange of a capital asset.

How do cryptocurrency markets look like in the coming days, weeks, or years?

It’s an exciting time and the everyone can invest safely and securely. The future is cryptocurrency.

Are the markets for different altcoins or largely independent affect the market cap for crypto?

Altcoins are considered meme coins and tweets and reddit do affect these coin and the market cap is endless.

How can I predict what will happen next?

Unfortunately, stock market is very hard to predict, but using the models that I created can help with a better outcome of crypto.

Conclusion

ARIMA performs well for next day’s predictions but performs poor for longer terms like given last few days price predict next 5-7 days prices. RNN perform consistently up to 6 days. Random Forest usually perform poorly on validation set and unseen patterns. RNN and ARIMA model perform well but fail to predict the hype or the unusual spike caused by rumors or fake news. One can use sentiments from twitter to incorporate these changes.

Reference:

cryptocurrency prices, charts and market capitalizations. CoinMarketCap. (n.d.).Retrieved October 3, 2021, from https://coinmarketcap.com/.2. Polygon.io. (n.d.). Retrieved October 3, 2021,

Sigalos, M. K. (2021, October 2). Ethereum had a rough September. here's why and howit's being fixed. CNBC. Retrieved October 3, 2021,romhttps://www.cnbc.com/2021/10/02/ethereum-had-a-rough-september-heres-why-and-how-it-gets-fixed.html.

Yahoo! (n.d.). The Crypto Daily – Movers and Shakers – october 2nd, 2021. Yahoo!Finance. Retrieved October 3, 2021, from https://finance.yahoo.com/news/crypto-daily-movers-shakers-october-002300815.html.6. Srk. (2021, July 7).

Cryptocurrency historical prices. Kaggle. Retrieved October 3, 2021,from https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory.
Taniaj. (2018, June 3). Cryptocurrency price forecasting. Kaggle. Retrieved October 3,2021, from https://www.kaggle.com/taniaj/cryptocurrency-price-forecasting.

Kash. (2021, September 11). Ethereum Cryptocurrency Historical Dataset. Kaggle.Retrieved October 3, 2021, from https://www.kaggle.com/kaushiksuresh147/ethereum-cryptocurrency-historical-dataset.

Sudalairajkumar. (2017, November 8). Cryptocurrency data pull. Kaggle. RetrievedOctober 3, 2021, from https://www.kaggle.com/sudalairajkumar/cryptocurrency-data-pull.

Coindesk: Bitcoin, Ethereum, crypto news and Price Data. CoinDesk Latest HeadlinesRSS. (n.d.). Retrieved October 3, 2021, from https://www.coindesk.com/learn/.