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

Intro: What is Ethereum?

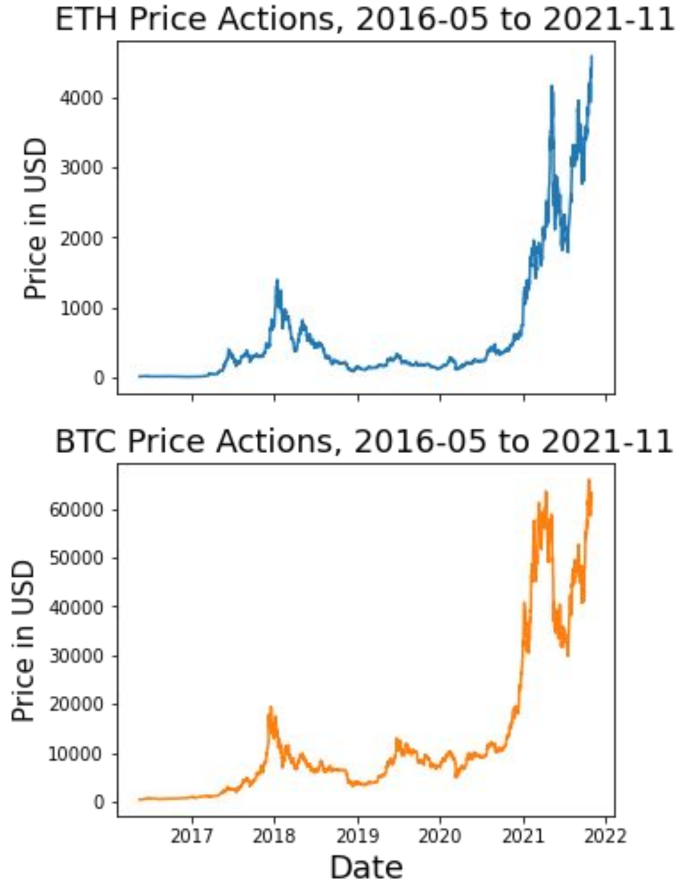
- World’s largest blockchain ecosystem
- Powered by its native cryptocurrency Ether (ETH)
- ETH is the 2nd largest cryptocurrency by market cap., behind Bitcoin (BTC)
- ETH is actively traded 24/7/365

Raw Datasets

- Reddit posts related to ETH ~ 1.1 million rows
- BTC daily price actions
    - Has fundamental differences with ETH
- Direct ETH competitors: ADA & SOL

Data Assembly

- Group the Reddit data by dates
- Match with the BTC price action data by dates
- Training set: from 2016-05 to 2019-10
- Testing set: from 2019-07 to 2021-07
- ‘Wild’ set: from 2021-07 to 2021-11



ETH Price Actions, 2016-05 to 2021-11

BTC Price Actions, 2016-05 to 2021-11

Spark Pipeline

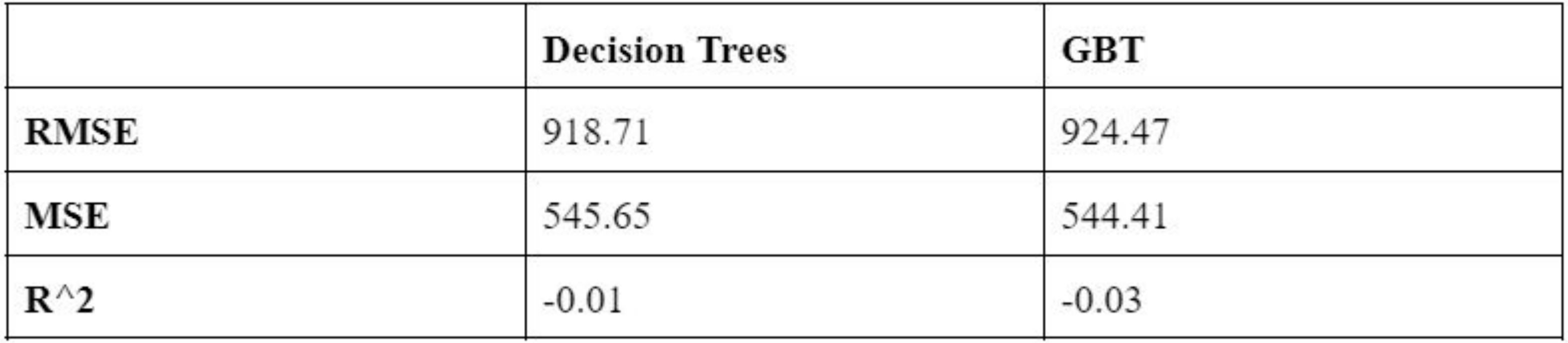
Four Stages:

1. Tokenizing the text data
2. Stopwords filtration
3. Word2Vec
4. Assemble

Two Models:

1. Decision Trees Regressor
2. Gradient-Boosted Trees (GBT) Regressor

Evaluations

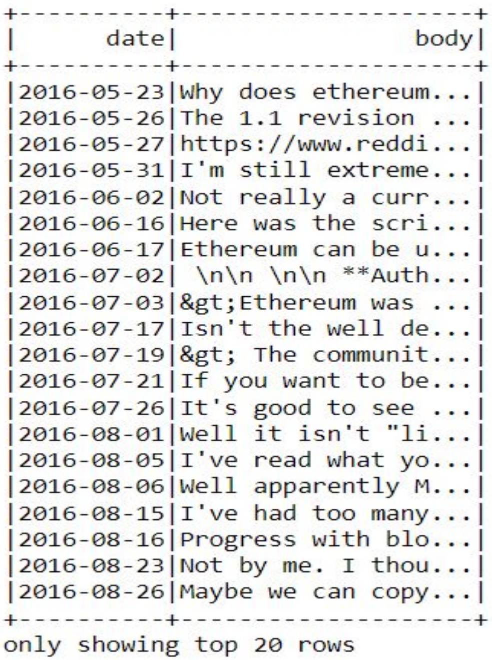


	Decision Trees	GBT
RMSE	918.71	924.47
MSE	545.65	544.41
R^2	-0.01	-0.03

Not Ideal !!!

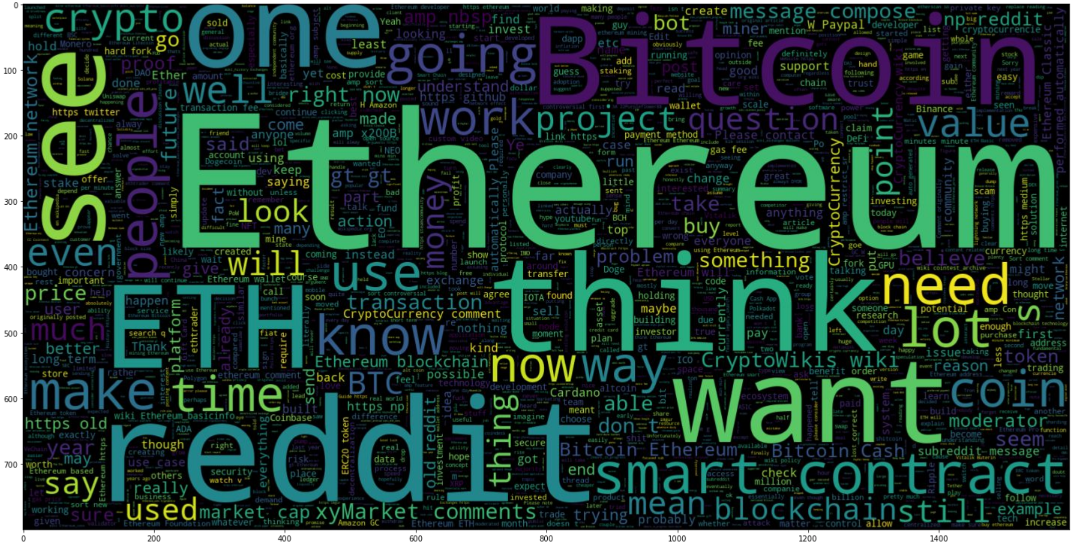
Reasons

1. The quality of the Data



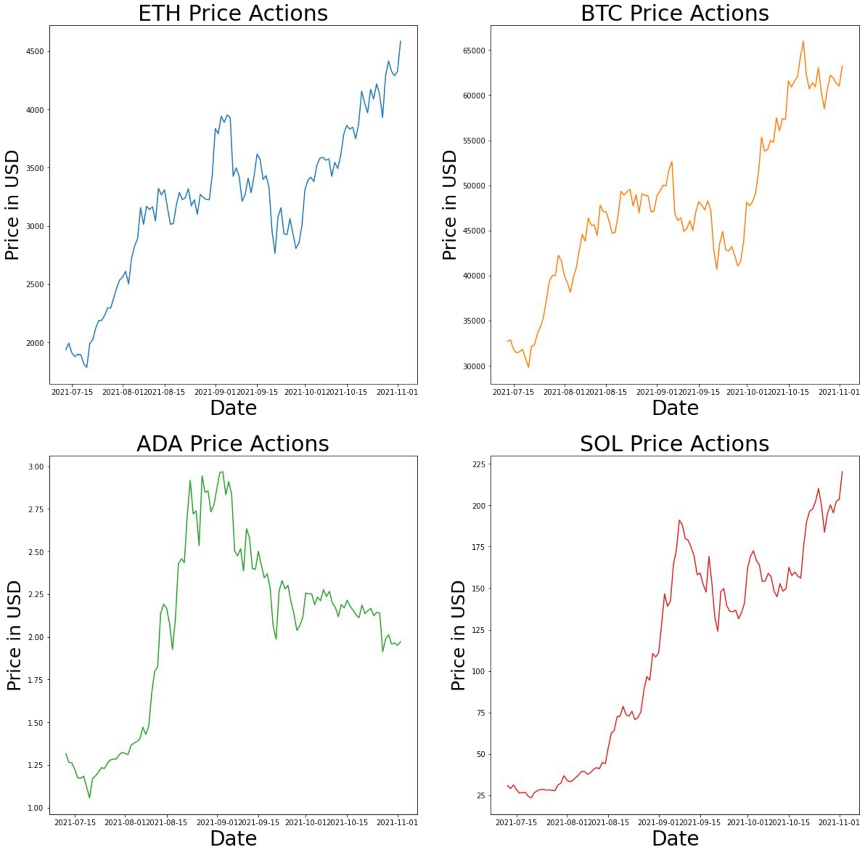
```
+----------+--------------------+
|      date|                body|
+----------+--------------------+
|2016-05-23|Why does ethereum...|
|2016-05-26|The 1.1 revision ...|
|2016-05-27|https://www.reddi...|
|2016-05-31|I'm still extreme...|
|2016-06-02|Not really a curr...|
|2016-06-16|Here was the scri...|
|2016-06-17|Ethereum can be u...|
|2016-07-02| \n\n \n\n **Auth...|
|2016-07-03|&gt;Ethereum was ...|
|2016-07-17|Isn't the well de...|
|2016-07-19|&gt; The communit...|
|2016-07-21|If you want to be...|
|2016-07-26|It's good to see ...|
|2016-08-01|Well it isn't "li...|
|2016-08-05|I've read what yo...|
|2016-08-06|Well apparently M...|
|2016-08-15|I've had too many...|
|2016-08-16|Progress with blo...|
|2016-08-23|Not by me. I thou...|
|2016-08-26|Maybe we can copy...|
+----------+--------------------+
only showing top 20 rows
```

Word Cloud for Ethereum Posts on Reddit



Reasons

1. The quality of the Data
2. Data features variety insufficiency
3. Cryptocurrency and blockchain are too young



Conclusion

- The Reddit data needs more processing before using it in the test-train process
- More features and varieties in the dataset are need, in order to build reliable price prediction models
  - Wait a few years for the army of ETH-killers to have enough data

Any questions?