Crypto Price Prediction: Design Document

Last Update: 6/28/19
First Proposed: 6/28/19
Author(s): Max Matuska, Courtney Luk, Lu Yu, Anthony Burre
Reviewer(s): dylanbaker@google.com
Status: [Draft]

Objective

- Predict cryptocurrency price performance
- Incorporate various factors into price predictions including but not limited to news article mentions, Twitter sentiment analysis, and number of trading application downloads
- Create a platform displaying our model's predictions in real-time

Background

super clear, awesome job here

Cryptocurrency is a digital currency that utilizes blockchain technology to keep records of transactions on a distributed, decentralized ledger. The blockchain operates through a network of many computers around the world which collectively verify and secure all transactions. The computing power required for this is garnered by compensating those who dedicate their processors to the task with additional units of whatever cryptocurrency they are helping keep track of. This is known as cryptocurrency mining.

A popular example of cryptocurrency is Bitcoin, a peer-to-peer digital currency that uses cryptography to manage transactions without any singular authoritative control. It was released in 2009 and was the first cryptocurrency based on blockchain technology. Due to its characteristics of decentralized management, anonymous transactions, limited supply, and transparent payment, it has become very popular in recent years.

Cryptocurrency prices, like stock prices, tend to be volatile and unpredictable. However, their prices are adjusted according to the difference between supply and demand. If we can acquire some measure of the demand for specific cryptocurrencies, we may be able to predict price trends to assist in future buy/sell decisions.

Overview

Overall notes on this section: Super clear writing! Given how many models you're combining, make sure you're clear about listing all the models you're using, broadly why you chose them, how you're combining them, and what the end user platform looks like. Then you could maybe include a subsection expanding on each of your models (are you training it yourself? why are you using X architecture? What pitfalls might you encounter? Has similar work been done?)

The first step of this project will be data collection and organization, as we are planning to incorporate multiple live data sources into our model's predictions. These sources include tweets from Twitter, Reddit threads/comments within subreddits that contain the word "Bitcoin", "investment", and "cryptocurrency", articles from reputable news sources that mention Bitcoin or

How? There are a lot of approaches. Are you building this model yourself or using a pretrained one?

cryptocurrency in general, and number of daily downloads of Bitcoin-related trading applications from various app stores. From here, we will conduct sentiment analysis on specifically tweets and Reddit comments to assign a numerical score that indicates how positive or negative people's current reactions are towards the Bitcoin market. Since the data we want to work with is live data, we can't only rely on historical data that has already been gathered by other sources such as Google Trends.

Are you also using this? ↗ If so, say so explicitly.

maybe give the acronyms here as well (in parentheses) because it's so widely-used

This project will use a long short-term memory neural network to predict movement within the cryptocurrency market. This network will take both live data from the cryptocurrency market and sentiment information from the various sources listed above to predict the future price of a particular cryptocurrency. The price of any cryptocurrency is influenced by many factors. Additionally, using the overall performance trend of other cryptocurrencies may also help to predict the future price of other cryptocurrencies depending on their correlation. Later changes to add additional forms of sentiment analysis input from other sources may be required if they're relevant and improve our models accuracy.

Also- why LSTM over other architectures? ⇓ can go in non-overview section

Because the price of Bitcoin is largely dependent on public excitement and anticipation of price increases, it tends to jump when news outlets publish articles about its growth. We may create a model that tries to predict the publishing of news articles relating to bitcoin and other major cryptocurrency prices. The input of this model could be historical data on the price fluctuations of crypto prior to articles addressing it. The output of this model could then be used to inform our main model.

could go in non-overview section

give some justification as to why this should be a good predictor- are you assuming the fluctuations are what's being reported on most of the time? If so, be explicit in this assumption. Maybe consider citing other things that might prompt articles to be written- somebody famous saying something, some notable purchase, new technological developments in blockchain, business decisions (e.g. Facebook's new currency) — and why, in light of this, fluctuations are still a good predictor.

Develop the platform side a bit more. Is this an app? Mobile-friendly website?

what panels are you including here, and how are they being populated?

Fine if you don't know yet / are just brainstorming, but given how much is going on here I'd want a little bit more description (even a few sentences)

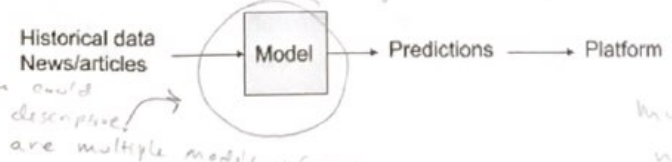


Diagram could be more descriptive: are multiple models informing a central model? Are results being combined in another way?

* remind me to show examples of model architecture diagrams

My gut says it's too noisy but I don't know this domain well.

Detail Design/Ideas

Detail write ups on design approaches, ideas/experiment planned, requirements, timeline/schedule, etc.

Tips:

- Focus on design, not implementation
- A picture is worth a thousand words. A diagram or drawing of the data and model at times is much easier to understand than words
- If describing alternatives, eg: approach, platform, algorithm - explain why it's not part of the design
- Do go back and update if during implementation you took a different approach that originally designed

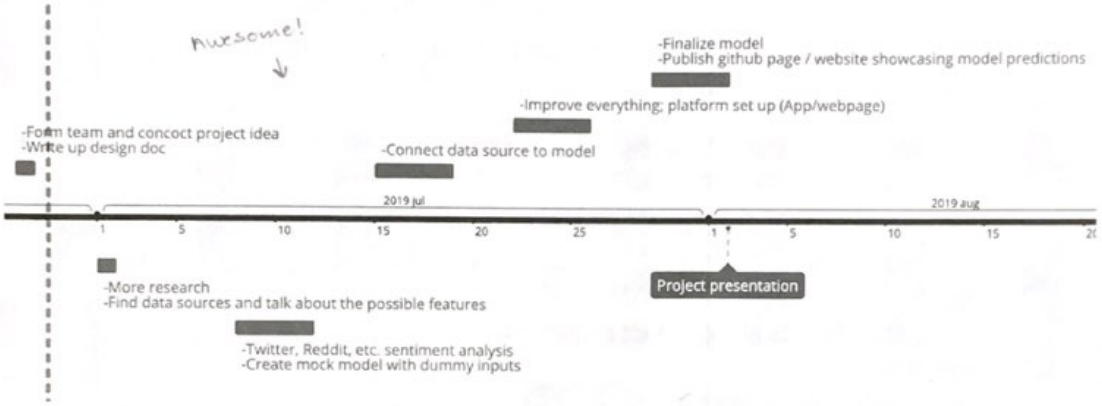
Data Sources:

1. Apple App Store API
    a. Download and rating numbers by day for coinbase, robinhood, and other crypto trading apps
2. Twitter
    a. Extract tweets and conduct sentiment analysis
3. Google News
    a. Other news conglomerations
    b. Ticker mentions
    c. API key: `72bc383e0bd44a5cb0997c468f7e0a2e`
4. Google Play Store
5. Google search
    a. Google search trends
6. Reddit
7. Crypto Price API
8. Blockchain API
    a. https://www.blockchain.com/api
9. Reuters

Preliminary Timeline:

DATE	TASK
Week 1 [6/27 - 6/28]	-Form team and concoct project idea -Write up design doc -Begin Twitter, Reddit, etc. sentiment analysis collection

Week 2 [7/1 - 7/2], Break	-Research -Find data sources and talk about the possible features
Week 3 [7/8 - 7/12]	-Create mock model with dummy inputs
Week 4 [7/15 - 7/19]	-Connect data source to model
Week 5 [7/22 - 7/26]	-Improve everything; platform set up (App/webpage) -Bug fixes ← usually don't need to say this explicitly, there are always bug fixes lol
Week 6 [7/29 - 8/2]	-Finalize model -Publish Github page / website showcasing model predictions



Fairness Considerations

Write a one-paragraph story describing a fictional person who was positively affected by a model trained with this data

Alex and Reese are siblings who have recently inherited a fortune, and are smart enough to diversify their investments. They realize the potential gains associated with cryptocurrency, and begin using our model on a daily basis to determine whether to invest or divest funds from different cryptocurrencies. With the help of our model, they manage to avoid losses on down days for crypto, and take full advantage of most days during which crypto makes strong gains. At the end of a month, they have made more in crypto than via traditional investments, and a news article gets published about their success. Our model takes this news article into account and correctly anticipates further crypto gains. As a result, Alex and Reese make even more money. Despite the gaping hole in their family life following the tragic passing of their parents, Alex and Reese have more money than they could ever need. lmao

→ Also worth considering:
- Reporters gaming the system to make $
- Twitter/Reddit bots gaming the systems
- Blind spot in training data causes huge losses

- not adequately conveying risks → over confidence in product → risky investing → $
- Security breach → hacker knows how people plan to invest → $
- people inventing cryptocurrencies to game the system?? It's wild out there!!

Write a one-paragraph story describing a fictional person who was negatively affected by a model trained with this data

Riley and Skylar have never met, but both their sets of parents give them both the same options: receive $200,000 upon turning 18, or have college paid for. They both decide on the 200 grand, as they believe with the right investments this can set them up for a lifetime of leisure. Riley has faith in the future of cryptocurrency, and buys $200,000 worth of bitcoin. Skylar finds our model and realizes that it usually predicts bitcoin price trends accurately, deciding to follow its suggestions by investing and divesting each day based on our predictions. One day, the entire crypto market plummets despite our model predicting an up day. Skylar and Riley, penniless and depressed, turn to substance abuse to numb the constant pain accompanying sober reflection on their actions. After two, separate, dizzying ten month benders, they meet in court ordered rehab, where they fall in love. They come to understand that true happiness is cultivated through human connection, not monetary success, and live happily ever after. we've all been there

Describe at least two sources of bias the particular model in your story could have

- known issue with Twitter: not classifying some dialects of English as English, could perform worse at sentiment analysis for some subpopulations.

Reporting Bias: One aspect of our model will be an analysis of how news articles influence cryptocurrency price. However, regardless of the number and variety of sources we pull articles from, we will miss large chunks of the internet, and therefore will not totally account for the sentiment of all potential cryptocurrency investors. → Are there groups of people who are left out more than others? (not always true, but usually)

Coverage Bias: Our model will be pulling posts on reddit and twitter pertaining to cryptocurrency in order to analyze public sentiment and predict price trends. However, we may not be able to train our model on historical data of this type, specifically that from the weeks leading up to the last cryptocurrency crash. This could leave our model unable to anticipate the worst loss days for crypto value, making it unhelpful for those trying to maximize gains before a bubble burst.

Describe at least one way we could modify the model to mitigate this bias

→ or otherwise adjust for/account for

Describe at least one way we could modify the dataset to mitigate this bias

We could spend extra time scraping news articles from a wide variety of news sources, using search engines other than Google to lessen the chance of missing significant contributors to public opinion. In this way our dataset will be more robust and we can mitigate reporting bias. what kinds of news sources are you worried you'll miss?

Describe at least one way we could modify the context surrounding the model to mitigate this bias

We can be transparent about the sources of our data ✓

References

Elliptic Curve Cryptography in Practice
Blockchain Explained