Is Bitcoin better than currency?

On December 14th, 2017 Bill Gates said “Bitcoin is better than currency”. I created a model to predict the price of Bitcoin using “internet buzz” to find out if Bill Gates (net worth 90 Billion USD) was right.

**Data Acquisition**

Live streaming the price of bitcoin can be accessed from all cryptocurrency exchange sites, this was the target of my models. I chose to get my data using an API at CryptoCompare.com based on the well written documentation and additional social media metrics they provided.

Along with Bitcoin’s price, I included various measurements from Facebook (Bitcoin’s total likes and total number of users talking about Bitcoin), Twitter (Bitcoin’s total followers, and total favorites), Reddit (Bitcoin’s subreddit “reddit.com/r/bitcoin” total subscribers, posts and comments per hour and day), GitHub (Stars, subscribers, open issues, last update, last push, and forks), and CryptoCompare Forums (Bitcoin’s followers, posts, comments, and page views). Full details in Appendix/Data Dictionary.

I created functions in Python using the libraries Requests, JSON, and Pandas to connect to CryptoCompare’s API every 15 seconds, download the raw data, and append it as a new row in a csv file along with a timestamp.

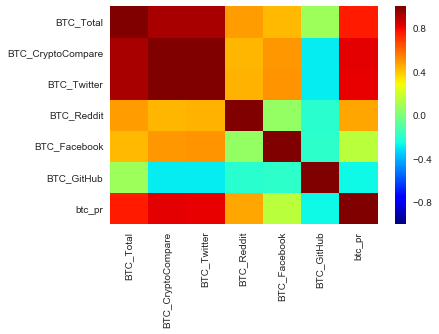
**Data cleaning, transformation**

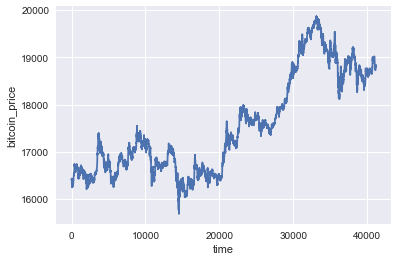
I prepared the 41,000 rows of data taken between the times of 2017\_1210\_23:13:35 and

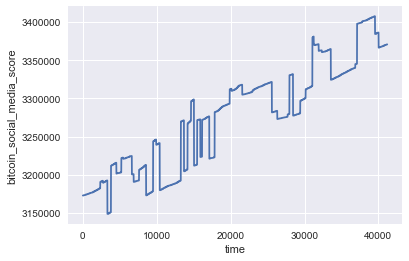
2017\_1218\_19:05:08 for modeling. I removed currency formatting from the coin price features (Dollar signs, commas, and decimals points). I aggregated individual site metrics into a total score for each website and labeled the rows.

**Visualization**

I used Matplotlib to visualize the data, looking for connections using a heat map correlation matrix and plots. I found GitHub appeared to have less correlation than I expected. Twitter and the CryptoCompare forums had the highest correlation.







**Modeling**

Due to the size of my dataset being over 10MB I used Databricks for modeling, a Scala cluster computing environment based on Apache Spark. This would allow me to create and test models much faster than on my laptop.

I imported the the csv file using sqlContext and set the headers. I set the bitcoin price as “y” and 5 features as ”X”. I split the data using randomSplit into train and test sets comprising 80% and 20% respectively. I created a Random Forest Regressor which combines many decision trees in order to reduce the risk of overfitting. I adjusted the number of trees by testing output scores from 30 to 100 trees in increments of 10, finding 35 as the most accurate, achieving an accuracy (r2) score of 0.9799777739508926.

**Results**

Given user activity metrics from around the internet I am able to predict Bitcoin’s price with 97.9% accuracy.

Limitations: It is impossible to predict the future.

**Conclusion**

The trend suggests that Bill Gates is correct, and I predict the price of Bitcoin to continue it’s rise in 2018.

**Next steps**

To improve the model I would apply a time series, additive model to the data such as Facebook Prophet to uncover “seasonality” and “holidays”, this will help pinpoint important events which can be added as features to increase model accuracy. Also, to improve server uptime and data integrity I would move the data collection script to an AWS cloud computer.

APPENDIX

Data Dictionary:

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter | Type | Always returned | Description |
| Response | string | yes | The type of the response (Success or Error) |
| Message | string | yes | The message for the response |
| Type | integer | yes | Integer representing the type of response. |
| Data | object | yes | Empty if there is no data to return or there is an error |
| Data:General | object | yes | The general social info |
| Data:General:Name | string | yes | The name of the exchange / symbol of the coin |
| Data:General:CoinName | string | yes | The name of the coin |
| Data:General:Type | string | yes | The type of the page: Webpagecoinp or Webpageexchangep |
| Data:General:Points | string | yes | The sum of all the individual points categories |
| Data:CryptoCompare | object | yes | The social data from our website |
| Data:CryptoCompare:SimilarItems | Array[5] | yes | Similar coins/ exchanges based on the number of total points |
| Data:CryptoCompare:SimilarItems[]:Id | int | yes | The id of the coin/exchange |
| Data:CryptoCompare:SimilarItems[]:Name | string | yes | The name of the exchange / symbol of the coin |
| Data:CryptoCompare:SimilarItems[]:FullName | string | yes | The full name of the exchange / name of the coin |
| Data:CryptoCompare:SimilarItems[]:ImageUrl | string | yes | The relative path (without https://www.cryptocompare.com) of the coin/exchange logo |
| Data:CryptoCompare:SimilarItems[]:Url | string | yes | The relative path (without https://www.cryptocompare.com) of the coin/exchange |
| Data:CryptoCompare:SimilarItems[]:FollowingType | int | yes | The type of following when someone wants to follow this coin/exchange (1 for coin, 2 for exchange) |
| Data:CryptoCompare:CryptopianFollowers | Array[11] | yes | Up to 11 followers (11 because in the front end if you are one of the 10 and you unfollow we still want to show 10) |
| Data:CryptoCompare:CryptopianFollowers[]:Id | int | yes | The id of the cryptopain that is following |
| Data:CryptoCompare:CryptopianFollowers[]:Name | string | yes | The name of the cryptopain that is following |
| Data:CryptoCompare:CryptopianFollowers[]:ImageUrl | string | yes | The avatar image relative path (without https://www.cryptocompare.com) of the cryptopain that is following |
| Data:CryptoCompare:CryptopianFollowers[]:Url | string | yes | The the profile page relative path (without https://www.cryptocompare.com) of the cryptopain that is following |
| Data:CryptoCompare:CryptopianFollowers[]:Type | string | yes | The type of the follower : Cryptopian |
| Data:CryptoCompare:Comments | int | yes | The total forum comments on the coin/exchange |
| Data:CryptoCompare:Points | int | yes | The total CryptoCompare points of the coin/exchange |
| Data:CryptoCompare:Posts | int | yes | The total forum posts on the coin/exchange |
| Data:CryptoCompare:Followers | int | yes | The total number of followers of the coin/exchange |
| Data:CryptoCompare:PageViewsSplit | obj | yes | The page views of the coin/exchange split by category |
| Data:CryptoCompare:PageViewsSplit:Overview | int | yes | The Overview page views of the coin/exchange |
| Data:CryptoCompare:PageViewsSplit:Markets | int | no | The Markets page views of the coin |
| Data:CryptoCompare:PageViewsSplit:Analysis | int | no | The Analysis page views of the coin |
| Data:CryptoCompare:PageViewsSplit:Charts | int | no | The Charts page views of the coin |
| Data:CryptoCompare:PageViewsSplit:Reviews | int | no | The Reviews page views of the exchange |
| Data:CryptoCompare:PageViewsSplit:Trades | int | yes | The Trades page views of the coin/exchange |
| Data:CryptoCompare:PageViewsSplit:Forum | int | yes | The Forum page views of the coin |
| Data:CryptoCompare:PageViewsSplit:Influence | int | yes | The Influence page views of the coin/exchange |
| Data:CryptoCompare:PageViews | int | yes | The total page views of the coin/exchange (sum of all PageViewsSplit) |
| Data:Twitter | obj | yes | The Twitter social data |
| Data:Twitter:followers | int | no | Total Twitter followers |
| Data:Twitter:lists | int | no | Total Twitter lists that the account is part of |
| Data:Twitter:favourites | int | no | Total Twitter favourites |
| Data:Twitter:statuses | int | no | Total Twitter statuses |
| Data:Twitter:account\_creation | ts | no | Twitter account creation date |
| Data:Twitter:name | string | no | Twitter account name |
| Data:Twitter:link | url | no | Twitter account link |
| Data:Twitter:Points | int | yes | Twitter total points |
| Data:Reddit | obj | yes | The Reddit social data |
| Data:Reddit:subscribers | int | no | Total Reddit subscribers |
| Data:Reddit:active\_users | int | no | Total Reddit active users |
| Data:Reddit:community\_creation | ts | no | Reddit community creation date |
| Data:Reddit:posts\_per\_hour | int | no | Total Reddit posts per hour |
| Data:Reddit:posts\_per\_day | int | no | Total Reddit posts per day |
| Data:Reddit:comments\_per\_hour | int | no | Total Reddit comments per hour |
| Data:Reddit:comments\_per\_day | int | no | Total Reddit comments per day |
| Data:Reddit:link | url | no | Reddit account link |
| Data:Reddit:name | string | no | Reddit account name |
| Data:Reddit:Points | int | yes | Reddit total points |
| Data:Facebook | obj | yes | The Facebook social data |
| Data:Facebook:likes | int | no | Total Facebook likes |
| Data:Facebook:is\_closed | bool | no | Is Facebook account closed |
| Data:Facebook:name | string | no | Facebook account name |
| Data:Facebook:talking\_about | int | no | Total number of Facebook users that are talking about this page |
| Data:Facebook:Points | int | no | Facebook total points |
| Data:CodeRepository | obj | yes | The code repository social data - only for certain coins |
| Data:CodeRepository:List | Array[] | no | List of code repositories for this coin |
| Data:CodeRepository:List[]:stars | int | no | Code repository stars |
| Data:CodeRepository:List[]:language | string | no | Code repository programming language (C++,Java,GO,Python,JavaScript etc) |
| Data:CodeRepository:List[]:forks | int | no | Code repository total forks, including forks of forks |
| Data:CodeRepository:List[]:open\_total\_issues | int | no | Code repository open total pull requests and bugs |
| Data:CodeRepository:List[]:subscribers | int | no | Code repository total subscribers |
| Data:CodeRepository:List[]:size | int | no | Code repository size in bytes |
| Data:CodeRepository:List[]:url | url | no | Code repository url |
| Data:CodeRepository:List[]:last\_update | ts | no | Code repository date of last update |
| Data:CodeRepository:List[]:last\_push | ts | no | Code repository date of last push |
| Data:CodeRepository:List[]:created\_at | ts | no | Code repository date of creation |
| Data:CodeRepository:List[]:fork | bool | no | Is code repository a fork (true/false) |
| Data:CodeRepository:List[]:Source | obj | no | If it is a fork, this has actual data about the coin it is a fork of - the source (parent of parent or higher - this is the same as the parent when this is the first fork). If we have the coin on crypto compare it contains data about the coin, otherwise it has the name and url of the code repo |
| Data:CodeRepository:List[]:Source:Name | bool | no | Name of the source code repo |
| Data:CodeRepository:List[]:Source:Url | url | no | Url of the source code repo |
| Data:CodeRepository:List[]:Source:InternalId | int | no | -1 if we don't have it, the actual id if we have it. |
| Data:CodeRepository:List[]:Source:InternalData | obj | no | Details about the coin page this is a fork of |
| Data:CodeRepository:List[]:Source:InternalData:Id | int | no | Id the coin page this is a fork of |
| Data:CodeRepository:List[]:Source:InternalData:Name | string | no | Name the coin page this is a fork of |
| Data:CodeRepository:List[]:Source:InternalData:Symbol | string | no | Symbol the coin page this is a fork of |
| Data:CodeRepository:List[]:Source:InternalData:ImageUrl | string | no | The logo image relative path (without https://www.cryptocompare.com) of the coin page this is a fork of |
| Data:CodeRepository:List[]:Source:InternalData:Type | string | no | The type of the page this is a fork of - Webpagecoinp |
| Data:CodeRepository:List[]:Parent | obj | no | If it is a fork, this has actual data about the coin it is a fork of - the direct parent. If we have the coin on crypto compare it contains data about the coin, otherwise it has the name and url of the code repo |
| Data:CodeRepository:List[]:Parent:Name | bool | no | Name of the parent code repo |
| Data:CodeRepository:List[]:Parent:Url | url | no | Url of the parent code repo |
| Data:CodeRepository:List[]:Parent:InternalId | int | no | -1 if we don't have it, the actual id if we have it. |
| Data:CodeRepository:List[]:Parent:InternalData | obj | no | Details about the coin page this is a fork of |
| Data:CodeRepository:List[]:Parent:InternalData:Id | int | no | Id the coin page this is a fork of |
| Data:CodeRepository:List[]:Parent:InternalData:Name | string | no | Name the coin page this is a fork of |
| Data:CodeRepository:List[]:Parent:InternalData:Symbol | string | no | Symbol the coin page this is a fork of |
| Data:CodeRepository:List[]:Parent:InternalData:ImageUrl | string | no | The logo image relative path (without https://www.cryptocompare.com) of the coin page this is a fork of |
| Data:CodeRepository:List[]:Parent:InternalData:Type | string | no | The type of the page this is a fork of - Webpagecoinp |
| Data:CodeRepository:List[]:open\_pull\_issues | int | no | Code repository number of open pulls |
| Data:CodeRepository:List[]:closed\_pull\_issues | int | no | Code repository number of closed pulls - either merged or deemed unsuitable |
| Data:CodeRepository:List[]:closed\_total\_issues | int | no | Code repository number of closed issues (pulls,bugs,features etc) |
| Data:CodeRepository:List[]:open\_issues | int | no | Code repository number of open bugs,features etc without pulls |
| Data:CodeRepository:List[]:closed\_issues | int | no | Code repository number of closed bugs,features etc without pulls |
| Data:CodeRepository:Points |  |  |  |