Sentiment Analysis for BitCoin

A demonstration of Data Science concepts

General Assembly: Data Science

Aug XX, 2015

Desiree Sylvester

Table of Contents

Problem Statement and Hypothesis

BitCoin is a form of digital currency that is not tied to any central organization or government entity. Part of its appeal is derived from features like encryption, exchange history, and ownership anonymity. Companies conducting business related to BitCoin -- such as currency exchanges, cloud storage, and exchange traded funds -- endorse the digital currency trend as the future of money[1].

The purpose of this paper is to evaluate two questions:

- Is public opinion changing over time to suggest that broader adoption of digital currency is possible?
- Is social media sentiment regarding BitCoin a leading or lagging indicator in its price?

These questions translate roughly into the following hypotheses:

Public Opinion

H_0 : There is no change in public opinion regarding BitCoin.

H_1: There is change in public opinion regarding BitCoin.

BitCoin Sentiment

H_0 : There is no relationship between StockTwit sentiment and BitCoin pricing.

H_1: There is a relationship between StockTwit sentiment and BitCoin pricing.

Data Sources

The datasets that will help test these hypotheses are collected from a number of sources. Coin Center is a non-profit research and advocacy group that promotes cryptocurrency-friendly regulatory climates, like BitCoin. It also conducts a monthly survey measuring American attitudes toward BitCoin, which it has been running since September, 2014. It publishes the data set for each of the last eight surveys on its website. The Coin Center survey data is valuable because questions germane to this topic are scalar variables, lending to statistical evaluation.

The survey instrument (see Appendix) contains ten questions gauging familiarity with BitCoin, the respondents' perspective of its trustworthiness and usefulness. Each month's dataset contains approximately one thousand responses.

[1] http://www.marketwatch.com/story/how-bitcoin-represents-the-future-of-money-and-global-finance-2015-01-29

I considered two different data sets to provide BitCoin sentiment. The first tracks BitCoin mentions on Twitter and the second tracks them on StockTwits. Considering the general nature of information exchanged on StockTwits, I felt is was a better measure of informed opinion regarding BitCoin. Both data sets are structured the same, with measures for bullishness and bearishness in tweets. The data set also provides the total number of mentions for a given day. These data sets are published by PsychSignal and distributed through Quandl.

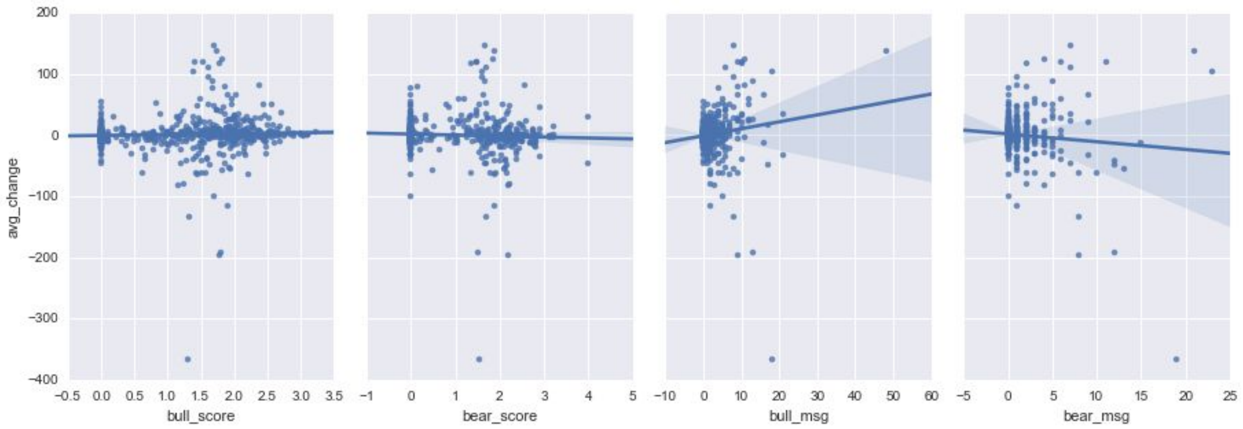
The last data set contains the average daily price of BitCoin in US dollars (USD) gathered across multiple exchanges. It is published by BitCoin Charts and distributed through Quandl. Rather than look at absolute price, I transformed it to show change in price instead.

Data Pre-processing

To prepare the StockTwits and BitCoin pricing data for analysis, I needed to merge the two dataframes using the StockTwits 'Date' field as the key. Scanning through the StockTwits data, each day did not have an entry for BitCoin mentions. By merging on the dates in the StockTwits dataframe, I get pricing information just for the days with mentions. Reviewing the column names, I noticed many of them have spaces. I updated the column names for the merged data to remove spaces.

Early Observations

I created a scatter plot of four features against the average change in price. I expected a stronger relationship between the intensity of bullish or bearish messages versus the number of messages, but was surprised to see a more pronounced slope in the latter.



[TODO: write segue into feature selection that discusses how scatter plot provides early clues about possible relationships.]

Feature Selection

TODO:

- create a boolean feature that evaluates the ‘avg_change’ value as True/False for positive or negative change, respectively

Model Selection

TODO:

- summarize reasons for selecting logistic regression
- evaluate the model against test set

Challenges and Successes

Business Application

Conclusion

References

Name	Link
Coin Center Survey	https://drive.google.com/open?id=0B-DuFJ-Q-6mEYjUzd2ZVVHBjTWM
BitCoin sentiment on StockTwits	https://www.quandl.com/data/PS1/BCOIN_ST-Bitcoin-BCOIN-Sentiment-Data-StockTwits
BitCoin average daily price	https://www.quandl.com/data/BAVERAGE/USD-USD-BITCOIN-Weighted-Price

Appendix

Coin Center: Survey Instrument

Question number	Question text
1	How familiar or unfamiliar are you with Bitcoin?
2	How frequently are you buying, receiving, or sending bitcoins, or have you never used Bitcoin?
3	Given what you know, how much do you *trust or distrust* Bitcoin?
4	Given what you know, how *useful or not useful* do you think Bitcoin is *as of today*?
5	Given what you know, how *useful or not useful* do you think Bitcoin will be *in the future*?
6	In a few words please describe the first thought or image that comes to mind when you think of Bitcoin.
7	Some things that make me distrust Bitcoin are:
8	Some things I think Bitcoin might be useful for are:
9	Some argue that Bitcoin is mostly used for legal transfers and investing; others argue Bitcoin is mostly used for crime, fraud, or extortion. Which comes closer to your view:
10	Governments still need to decide how to regulate Bitcoin. Some argue it should be banned, others want it left alone, some are in the middle. What's your view?

StockTwits Dataset Column Descriptions

Column	Meaning
Date	This is the date of the analyzed data.
Bullish Intensity	Our algorithms score each message's language for the strength of bullishness present on a 0-4 scale. 0 indicates no bullish sentiment measured, 4 indicates strongest bullish sentiment measured. 4 is rare.
Bearish Intensity	Our algorithms score each message for the strength of bearishness present in the message on a 0-4 scale. 0 indicates no bearish sentiment measured, 4 indicates strongest bearish sentiment measured. 4 is rare.
Bull - Bear	This indicator simply subtracts bearish_intensity from bullish_intensity to provide an immediate net score.
Bullish Messages	This indicator is the total count of bullish sentiment messages scored by the algorithm.
Bearish Messages	This indicator is the total count of bearish sentiment messages scored by the algorithm.
Total Messages	This indicator is the number of messages coming through our source data feeds and attributable to a symbol regardless of whether our sentiment engine can score them for bullish or bearish intensity.