Bitcoin Market Drivers: A Predictive Analysis

CIS 519 Danica Fine, Emily Jiang 20 November 2017

Checkpoints

1. Acquire data from various sources for the interval from approximately January 2012 to September 2017.
2. Cleanse Bitcoin data to reflect normal market patterns, i.e. remove data corresponding to holidays and weekends.
3. Transform minute-by-minute Bitcoin exchange data to a daily granularity.
4. Construct additional features for the Bitcoin data, e.g. overall market increase or decrease compared to previous day.

Summary of Work

As reflected in the above section, the majority of our work so far has involved acquiring, cleansing, and aligning the data from a variety of sources. (Please see the Data section for a summary of the data being utilized in this project.)

Cleansing and massaging data proved to be the most time-consuming part of the project up to this point. Fortunately, the datasets acquired from Bloomberg and Wharton were exceptionally clean, so we focused on removal of superfluous data points based on our financial knowledge. For example, we chose to use only the field "Level of the S&P 500 Index" from the SPX data, as it is known to accurately reflect equity market performance.

Bitcoin data required further manipulations. Since Bitcoin exchanges are always open, we first had to align our dataset with normal market data by removing data points corresponding to US holidays or weekends. The most computationally complex aspect of our work involved grouping the minute-level Bitcoin data by day and computing corresponding daily values, e.g. open and close prices are taken from the minimum and maximum raw timestamps from the original dataset, highs and lows are minimum and maximum values for the day, and volumes and weighted prices were taken to be the averages for that day.

From the cleansed Bitcoin data, additional features were derived. Namely, we created flags to reflect whether each of the pricing and volume fields increased or decreased when compared to the previous day's values. We also captured raw deltas between days.

Methodologies

The data manipulation aspect of the project has allowed us to become more familiar with a variety of Python libraries that we had not previously had the ability to utilize. To make the removal of holidays data easier, we acquired and made use of an extensive `holidays` library (we found that `datetime` provided enough functionality to remove weekends).

One of the more interesting pieces of the project has been our use of `pandas`, specifically for its group-by capabilities within the `DataFrame` module. Before importing this library, we attempted to make use of raw Python and `numpy` functions to group the minute data by day, but the script was objectively too verbose. `pandas` allowed us to succinctly group the data while making use of a number of aggregate functions, reminiscent of a basic SQL call.

Bitcoin Market Drivers: A Predictive Analysis

CIS 519 Danica Fine, Emily Jiang 20 November 2017

Next Steps

With the data appropriately cleansed and additional features derived, we now plan to delve into the machine learning aspect of the project. As our proposal suggested, we will begin by utilizing basic machine learning techniques and build out more complex classifiers as we increase our understanding of the available data and its features.

To start, we will conduct a survey of classification models where the models will attempt to classify whether data suggests an increase or decrease of Bitcoin's valuation for that day. As of now, we anticipate using decision trees, k-nearest neighbors classifiers, and support vector machines; these models are simple and their performance will provide insight into the data and, perhaps, prompt us to revisit and potentially improve the derived features. From there, we will delve into more complex models, including neural networks – both in `sklearn` and TensorFlow. However, we admit that, since our feature space is relatively limited, we are uncertain as to how a neural network will perform on the dataset.

As an additional exercise, we will then build out various regression models. Regression is not our primary focal point due to the fact that Bitcoin's overall valuation has changed so drastically even within such a limited time period; as such, we are reluctant to rely on a regression model that would attempt to convey such drastic changes. Instead, we feel that addressing the classification problem will provide useful insights regarding our feature space and hopefully guide us in building out regression models.

Data

This project was inspired by a dataset on historical bitcoin prices.[1] The dataset includes prices gathered from various exchanges at the minute-level, ranging from 2012 to 2017. Bitcoin pricing is the focal point of this project, but a number of other financial datasets will be vital to the model as a whole. We make use of the following:

- S&P 500 Index[2]: Daily closing prices
- VIX (CBOE Volatility Index)[2]: Daily closing prices
- FX (Forex) Top Currency Trading Cyptos[3]: China (CNY), Japan (JYP), Euro, and British Pound (benchmarked against USD)
- Interest Rates[3]: Federal rate, treasury short- and long-term, BBB corporate bond
- Treasury Prices[3]

[1] Historical Bitcoin Data, https://www.kaggle.com/mczielinski/bitcoin-historical-data/data

[2] Sourced from Bloomberg Desktop API

[3] Provided by Wharton Research Data Services, https://wrds-web.wharton.upenn.edu/wrds/