Title of your project report

Sean Maloney* Sophia Margareta Carayannopoulos†

May 9, 2018

Abstract

This is the abstract of your report ...

*Department of Statistics. Email: maloney4@wisc.edu
†Department of Statistics. Email: carayannopou@wisc.edu

1 Introduction

Recently, cryptocurrencies have started to appear all over social and mainstream media. While the technology behind them is interesting what interests most people is the potential money. With bitcoin increasing over 10x in value during 2017 and the entire cryptocurrency ecosystem being measured in the hundreds of billions malicious actors might be trying to game the system through manipulated social media for their own financial gain.

One of the main platforms used for cryptocurrency discussions is Reddit. The subreddit /r/CryptoCurrency/ has over 600k worldwide members and is specifically committed to the discussion and distribution of information about cryptocurrencies.

Our goal to use the submissions and comments within /r/CryptoCurrency/ as metric of media hype to see if we could predict a cryptocurrency day to day performance .

2 Gathering Data

Our data set from /r/CryptoCurrency/ was collected from January 1, 2017 to December 31, 2017 and included all posts and comments. The dataset in total contains 5 million comments, 132k posts and 22k active users, active users being all unique accounts that posted or commented. We investigation will consider the top 10 coins (excluding bitcoin) by market cap on January 1, 2017 which from largest to smallest are as follows: Ethereum, Ripple, Litecoin, Monero, Ethereum Classic, Dash, MaidSafeCoin, Augar, Steem, NEM.

Bitcoin is purposely excluded in our investigation as it is used as a baseline price unit. Each of the top 10 coin?s prices will be considered in bitcoin value. Pricing in terms of bitcoin rather then USD is important as all of the coins we consider are traded directly to bitcoin rather then USD pairs.

3 The Data Set

Our data set contained 26 predictor variables computed from the raw post and comment text. Each variable is computed over a single day. Each statistic for a coin has a corresponding control statistic for the day. This is im-

portant as the total number of posts and comments increased significantly throughout the year, see Figure XXXX.

The control prediction and coin prediction variables are listed and described (when unclear) below:

3.1 Control Variables

All statistics are computed for over a single day for all posts and comments.

1. Total Number Comments
2. Total Mean Comment Length

   Comment raw text length in characters
3. Total Comment Different Authors

   Total unique authors for both comments and posts based on username
4. Total Mean Comment Score Per Comment

   The mean comment score corresponds to the mean of the sum of upvotes and downvotes where upvotes and downvotes all have equal weight
5. Total Comment Number Fomo

   Number of times ?Fomo? or ?fear of missing out? appears as a literal in raw comment text
6. Total Comment Number Fud

   Number of times ?FUD? or ?fear, uncertainty and doubt? appears as a literal in raw comment text
7. Total Number Posts

   Total number of posts in /r/CryptoCurrency
8. Total Post Mean Score

   The mean post score corresponds to the mean of the sum of upvotes and downvotes where upvotes and downvotes all have equal weight
9. Total Posts Mean Length

   Mean text length of all posts in characters

10. Total Posts Number Different Authors

    Number of unique authors over all posts

11. Total Posts Mean Number Of Comments

    Mean number of comments posts receive

12. Total Posts Number Fomo

    Number of times “Fomo” or “fear of missing out” appears as a literal in raw posts text

13. Total Posts Number “Fud”

    Number of times “FUD” or ?fear, uncertainty and doubt? appears as a literal in raw posts text

3.2 Coin Specific Variables

For the coin specific variables the post or comment was determined to be “for” the coin if they contained the coins name or the coins trading abbreviation in their raw text. All statistics are computed for each coin over a single day in the same fashion are the control variables. They are listed below and follow the same descriptions are the control variables above:

1. Total Number Coin Comments
2. Total Mean Coin Comment Length
3. Total Number Coin Comments Different Authors
4. Total Mean Coin Comment Score
5. Total Coin Comment Number Fomo
6. Total Coin Comment Number Fud
7. Total Coin Number Submissions
8. Total Coin Mean Submissions Score
9. Total Coin Mean Submissions Length
10. Total Coin Number Submissions Different Authors
11. Total Coin Mean Submissions Number Of Comments
12. Total Coin Submissions Number Fomo
13. Total Coin Submissions Number Fud

4 Preliminary Look

Our first step was to look at the bitcoin prices of each coin over the year. We saw that there were three possibly distinct price groups. XXXPLOTS of all of them Upon further inspection, it seemed that that one distinct price group was: Augur, Steem, Ethereum, Ethereum Classic, which we refer to as Group 1. XXXPlots of group 1 They each have a very easy to follow distribution, unimodal, and bell shaped. It seemed that those in Group 1 had a peak between days 100 and 200. Upon further inspection it seemed that they also shared a peak around days 160 to 180.

The next distinct price group, Group 2, seemed to be: Dash, Litecoin, Monero, Maidsafe. These coins did not have an easy to follow curve like coins from Group 1. They seemed multimodal, but they also seemed to share a specific peak between 200 and 300. XXXPlots of group 2 They all shared a 20 day peak as well between.

There is finally Group 3, though they were not as complicated to follow as the second group, their trend was not like Group 1 and deserved to be a separate and distinct group. It is possible that this group is bimodal but they share a peak between day 100-200. When we looked closer we saw that this peak was between 120-150, leading us to further believe that this is a distinct group. XXXPlots of Group 3

We could not discern why these coins fell into distinct groups based on the merits of the coins. For example Ethereum and Steem are in the same group. Ethereum is still one of the top traded coins today. Steem is more of a niche coin and is not widely used, yet they fall into the same group. There may be a confounding variable that would be able to help more fully explain the similar trends. It is also possible that they are just random.

5 Statistical Analysis

When we had set out to do this project we knew we wanted to do some kind of prediction of the value of the coin. We knew that with all of this information we would be unable to do a straightforward linear regression of any sort. Our first step was that we had to make all the variables more manageable. The next step was looking at the odds of whether or not the coin would rise in value. Finally we wanted to see how robust our model was. We felt that it would appropriate to build four models. One model would

be using all the coins and the other three would be for each distinct group of coins. We wanted to see if separating them into groups would greatly improve the model, which would mean overfitting. Another thing we were looking for was to see if a certain group was making the model perform well while the other one or two were performing poorly.

As mentioned above, we have a lot of variables that could be taken into account. They variables can also be correlated. It seemed that the most appropriate step would be to perform Principal Component Analysis, PCA. This technique is used for data sets like ours where we need to take the variables into a lower dimension. There are certain assumptions that need to be met before performing PCA. We are assuming that the variables have a linear relationship since PCA will be unable to detect nonlinear subspaces. An issue with that is we are not sure this holds for all of our variables, but we will hold the assumption that they are. Another assumption that is made before performing PCA is that we have a large and representative sample. There is no reason to believe that our dataset is not representative, let alone comprehensive. Once we performed PCA on all the variables except Coin, Close, High, Low, Open, we saw that to explain 80% of the variation we need 5 variables. On closer inspection of the components, we could not find any of the variables more significantly weighted than others. This is another drawback of PCA, when compacting the data into a lower dimension you lose some interpretation and understanding of the data.

Our next step was to use those components to fit a logistic regression model. We thought it would be interesting to see if we would be able to predict the odds if a coin went up or not. This made our logistic linear regression model a binomial one. One of the assumptions that could possibly be problematic is that we assume that all the variables are independent. Though we used PCA to deal with collinearity, it does not take away that a comment might influence an up vote and so on and so forth. We are working on the assumption that everyone can think for themselves and is not influenced by one another. There is the assumption, once again, that we need a large data set: we have met this. Due to using the component in our logistic regression, we did not feel it was appropriate to formal equation. For those who would like to see the model with the components fitted, the R code is included in the appendix

After fitting the 4 models, we wanted to create confidence intervals for how often we are correct when the price goes up, when the price goes down, type I error, type II error. We decided that nonparametric bootstrap would

be appropriate since we do not know that distribution of the population we are sampling from. For the assumption of bias, we might have potentially run into an issue. This is because we are testing the model with the same data we used to fit the model. Another assumption that is made is that is that the data set is large enough: there is no reason to believe that the data set is not large enough. When we created our confidence intervals for the above parameters we found that in general out model did not perform very well and had high type II errors. XXXtables here. It seemed that Group 1 is the only group that performs decently but even then, it would not be enough that it would be likely that someone could be exploiting reddit for gain. There is also negative values in our confidence intervals. As we know it is impossible to get a negative amount correct or incorrect, but this shows how much variance is in our model.

Appendix: R code

```
m <- 10000
theta.hat <- se <- numeric(5)
g <- function(x) {
  exp(-x - log(1+x^2)) * (x > 0) * (x < 1)
}

x <- runif(m)     #using f0
fg <- g(x)
theta.hat[1] <- mean(fg)
se[1] <- sd(fg)
```