Social Media Predictors of Cryptocurrency Prices

Stat471 – Final Project

Arpan Bagui, Daniel Shi

December 18, 2021

Code to reproduce this report can be found on this GitHub link.

Table of Contents

Executive Summary

Cryptocurrency has proven to be a viable alternative to the physical banking system and has become an exciting investment over the past decade. Despite their relatively short histories, cryptocurrency prices have enjoyed euphoric heights, only to suffer severe drawbacks. In particular, social media has proven to be particularly important in the adoption of different cryptocurrencies. Several high profile incidents have produced drastic shifts in asset prices, and the resulting volatility has shaken investor confidence.

Despite this, over the past year, cryptocurrencies have begun to draw more attention from not only avid crypto fans, but also large institutional investors such as BlackRock and Renaissance Technologies. As such, being able to accurately predict cryptocurrency price movements will be increasingly valuable, and as students interested in investing, our final project examines the relationship between social media engagement over particular cryptocurrencies and their respective prices.

Our data observes social media metrics on various platforms from March 9 to April 28 of 2021. Notably, this is a rather small sample size to assess the true relationship between social media and the price of cryptocurrency. Despite this, we believe our data and subsequent analyses provides a useful window into a particularly turbulent period of cryptocurrency activity. Our dataset was directly scraped from three social media platforms: Reddit, Twitter, and Youtube. Explanatory variables varied from platform to platform, as each has their own unique engagement metrics (e.g., Reddit's upvotes, Twitter's retweets, Youtube's subscribers). These metrics were observed against our primary response variable: price, which was reported on the various cryptocurrency exchanges.

Our data was collected from Reddit, Twitter, and YouTube, with accompanying coin price information gathered from Yahoo! Finance. To turn our data into usable and comparable metrics, averages for each of the raw metrics were taken, before being put into our cleaned data table. We explored 4 models to try to find an optimal model that could most closely monitor coin price: ordinary least squares regression, ridge regression, LASSO regression, random forest, and boosting. Ordinary least squares regression had the lowest test error, but it was still high with a RMSE of 2328.

Our primary takeaway from the exploration was that extreme heterogeneity exists in social media responses based on particular cryptocurrencies. Despite the relative youth of the crypto market, each coin has become highly specialized and subject to emotive reactions on social media from particular groups of people. Despite our ordinary least squares model having the lowest MSE, it was still relatively high, especially when considering the ranges and scales of the prices of different coins. When assessing the project's goal of considering whether or not social media is a useful heuristic for understanding crypto price, we think that the answer is not that simple, and should be evaluated on a much more individual basis.

Introduction

Cryptocurrencies are still early in their adoption, which has resulted in large speculative shifts in their prices. Several bad actors during the early conception of cryptocurrency have shaken investor confidence. The bankruptcy of cryptocurrency exchanges Mt. Gox and Yapian Youbit in 2014 and 2017 have shown the potential downsides of the currency, and the FBI shutdown of the Silk Road marketplace due to the use of the exchange in the drug trade has shown the downsides of little oversight. Even on a normal basis, cryptocurrencies exhibit high volatility. In March 12 of 2021, Bitcoin peaked at $61,000, achieving a 300% annualized return for the past 10 years, only to drop 30% in the following month, and recover 12% the month after.

Beyond the growing pains of infancy, social media in particular has had a striking influence on the price of cryptocurrencies. On June 14 of 2021, Tesla founder Elon Musk tweeted the company would no longer be accepting the use of Bitcoin as payment due to environmental concerns, and the price of the coin dropped 15% almost immediately. Positively, social media has also increased adoption rates, especially through video platforms like YouTube. The complex nature of the blockchain technologies used to power cryptocurrencies represent a large barrier to adoption, but educational videos have had a positive impact on informing investors about the inner workings of cryptocurrency.

The goals of this study are to quantify the relationship between social media and cryptocurrency prices, identify which metrics are the most useful in predicting prices, and create useful models to do so. Unlike traditional assets like stocks, bonds, preferred shares, and ETFs that are constantly being moved between companies and investors, cryptocurrency is almost purely a consumer oriented currency at this stage, meaning social factors have a particularly large effect on its price. All of the key metrics across Reddit, Twitter, and Youtube will be used to measure their effects on the only response variable: price. Little academic work has been purely dedicated to this relationship, as the boom is relatively recent and Elon Musk's surprising influence on the currency has only been observed in the past year. We are interested in finding out if a particular platform (and perhaps a particular *type* of platform) has the strongest impact on cryptocurrency prices, and which indicators on those platforms are most important in determining those prices.

For institutional investors, the potential upside in cryptocurrency has generated significant interest, but the currency's inherent volatility and complexity has deterred widespread adoption away from other traditional assets. Younger investors and crypto only strategies are not only beholden to the peaks and troughs of cryptocurrency prices, but also lack a stable asset base from which to build out their investments. Without a significant systematic or infrastructural departure from its past, the recent crypto boom is at risk of falling to a similar fate as its past occurrences, resulting in large losses for speculative investors. Because there are so few underlying factors that can help predict movements in cryptocurrency prices, analyzing the effect that social media has on it is one of the only quantifiable ways to understand sentiment on cryptocurrencies. We hope that this exploration will provide some initial clarity on how social media can have a tangible and measurable effect on the rising cryptocurrency markets.

Data Description

Data Sources

Our data was directly scraped from the three target social media platforms: Reddit, Twitter, and Youtube. Cryptocurrency price data was gathered from Yahoo! Finance. All explanatory and response variable data was collected between the dates of March 9 to April 28 of 2021. Data collection from each of the three platforms was slightly different. The processes are listed in the Data Description section. We utilized the Reddit and Twitter API in addition to a chrome driver that automated YouTube data collection. We picked twelve coins with relatively large market capitalizations to collect data on. For Reddit, everyday from March 9th to April 28th, 2021, the top 500 posts from each of the twelve coins' subreddits were scraped from the API. For Twitter, every day from March 9th to April 28th, 2021, the top 10 posts regarding each of the twelve coins were scraped from the API. For YouTube, every day from March 9th to April 28th, 2021, the top 10 videos regarding each of the twelve coins were scraped using Selenium.

Data Cleaning Process

Our data cleaning process was guided by our goal of having one observation for each coin per day. For Reddit, we averaged subscribers, number of comments, score, and upvote ratio and counted how many unique users made posts across the 500 posts per day. For Twitter, we averaged retweets, favorites, account followers and counted how many unique users tweeted across the 10 posts per day. For Youtube, we averaged the channel subscribers, views, number of comments, number of comments and counted how many unique channels posted videos across the 10 videos per day. All of these aggregations were done through the SQL pull, and thus the raw data before the pull (on the AWS server) was not included in our repository. The raw data in our repository represents the data following the cleaning above, after the SQL pull.

Following this step, we had three uncleaned datasets for Reddit, Twitter, and Youtube that required an additional feature for the price of the coin each day and then needed to be merged together. For each uncleaned dataset, we added price information on the coin by matching the coin to its price using the Yahoo Finance API. We then merged the datasets together on the closing price column which was the same for each dataset. We then reordered and renamed the columns to make the dataset more comprehensible.

Data Description

Overall Observations

The data includes 544 observations, each with 17 features. The 544 observations are split across 11 monitored cryptocurrencies, with each observation representing every social media metric of that coin on a particular date across the observation period. In theory, there should be 51 observations for every coin, representing every day in the observation period. However, the scraper wasn't able to get values for each explanatory variable on every day for every coin. As such, some coins are slightly overrepresented and others are underrepresented in the data, as we chose to drop all the observations with incomplete observations.

Response Variable

Price (measured in USD) is the response variable. Each coin is traded on a cryptocurrency exchange, with theoretically infinite upsides. As such, price is a continuous variable measured to the cent. For investors and casual observers, price per coin is the most accessible proxy for understanding the value of the digital currency, as it isn't based on a real asset.

Explanatory Variables

For each of the 11 coins, there are 15 explanatory variables total from three social media platforms: Reddit, Twitter, and Youtube. For all 16 total explanatory variables, unless noted otherwise, the variables are continuous. They are listed below:

Coin Type: cryptocurrencies come in different varieties, and each utilize different technologies. Each is traded under their own name with different prices. The 11 coins we observed are:Bitcoin, Ethereum, Cardano, Ripple, Algorand, Dogecoin, Chainlink, Monero, Filecoin, Tezos, and Stellar. This variable is categorical

Reddit: Reddit is a social news website and forum where content is socially curated and promoted by site members through voting. Topics are split up via different forums (subreddits) where users may individually subscribe to.

- *Subscribers:* users may subscribe to a subreddit to receive recurring updates on a topic. We measured the total subscriber count for the dedicated subreddit for each coin.
- *Comments:* users may comment on posts. We measured the average number of comments for the top 500 posts within each subreddit
- *Upvotes*: users may upvote a post to express support. We measured the average number of upvotes for the top 500 posts within each subreddit
- *Upvote Ratio:* users may also downvote posts. We measured the average upvote to downvote ratio for the top 500 posts within each subreddit

- Unique Posts: heterogeneity exists in the users whose posts are in the top 500 when searched. We measured how many unique users have posts within the top 500 when a coin is searched (maxed at 500).

Twitter: Twitter is an online news and social networking site where people communicate in short messages called tweets. Each account is linked with its own feed, where creators and influencers essentially microblog original content, and users may choose to follow accounts who's content they like. Users may also engage with each other via replies, subtweets, and likes.

- *Retweets:* users may retweet tweets which will share that tweet with the retweeter's followers. We measured the average retweets for top 10 tweets when a coin is searched.
- *Favorites:* users may favorite tweets, which expresses interests/support in the tweet. We measured average favorited tweets for top 10 tweets when a coin is searched
- *Followers:* each user account can be followed, which alerts followers when they tweet something new. We measured average followers of accounts that posted tweets in the top 10 tweets when a coin is searched.
- *Unique Users:* heterogeneity exists in the users that post tweets within the top 10 when searched. We measured how many unique users have posts within the top 10 when a coin is searched (maxed at 10).

YouTube: YouTube is a video sharing service where users can watch, like, share, comment and upload their own videos. Content creators each have their own channels where to create video content, and users may choose to subscribe to individual content creators.

- *Subscribers:* users may choose to subscribe to individual channels to receive updates when they produce new videos. We measured the average number of subscribers for channels that produced videos within the top 10 when a coin is searched.
- *Views:* every time a user watches a video, a view is tracked. We measured the average views for the top 10 videos when a coin is searched.
- *Comments:* users may choose to leave comments on a video. We measured the average comments for the top 10 videos when a coin is searched.
- *Likes:* users may choose to like videos, expressing their support. We measured the average likes for the top 10 videos when a coin is searched.
- *Unique Channels:* heterogeneity exists in the channels that upload videos within the top 10 when searched. We measured how many unique channels have videos within the top 10 when a coin is searched (maxed at 10).

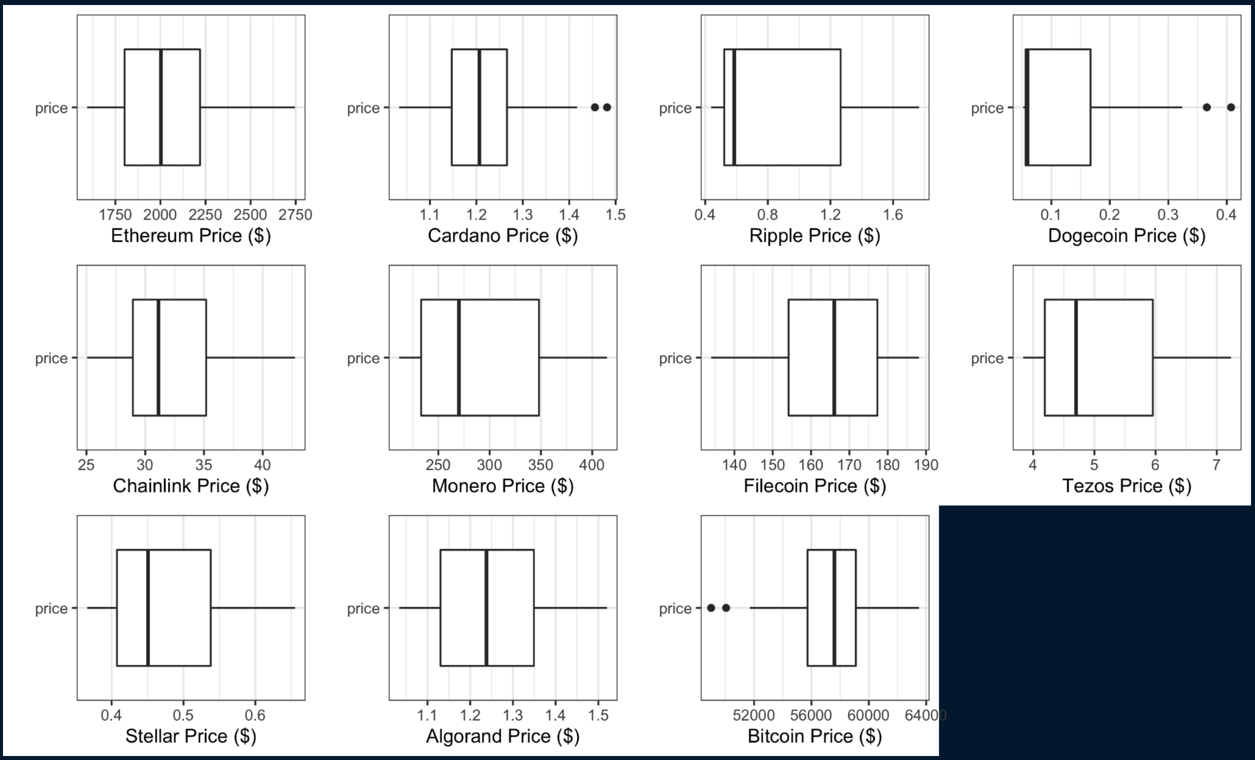
Data Allocation

Training and testing data were split on a roughly 80-20 split for each coin. In order to measure the aggregate effect of our observed social media indicators on cryptocurrency broadly, it made more sense to split the data so that each coin was represented in both the training and testing data. We tried our best to adhere to the 80-20 split, but in practice the split was 78-22 with 432 observations for training and 102 observations for testing.

Data Exploration

Response Variable

We began analyzing our data by looking at the distribution of coin prices in our training data. The results are summarized below in box-and-whisker plots for each of the coins:

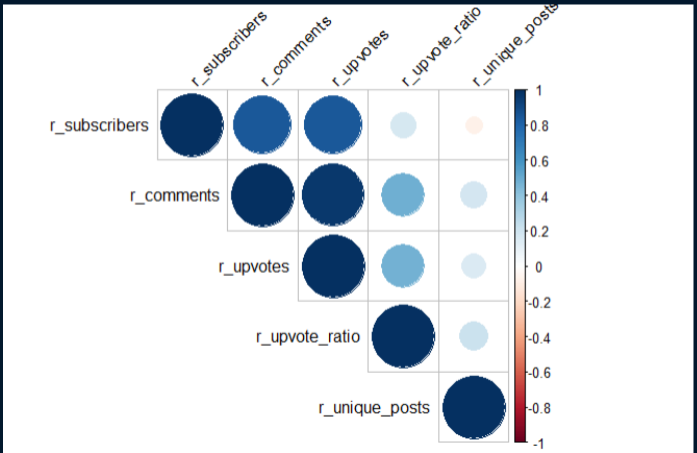


The median prices of the coins are as observed:$58,000/Bitcoin, $1.24/Algorand, $2,040/Ethereum, $1.20/Cardano, $0.55/Ripple, $0.01/Dogecoin, $31.5/Chainlink, $255/Monero, $165/Filecoin, $4.80/Tezos, $0.42/Stellar. As shown, there are obviously large differences in the coin prices, based on the actual coin itself, which will give coin type a relatively large magnitude in our regression models.

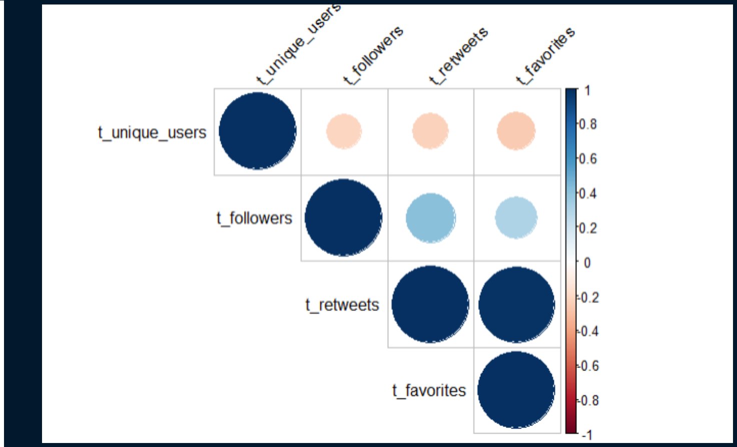
Explanatory Variables

To understand the correlations between each of our predictor variables, we began by examining the inter-platform metrics' correlation with each other, as certain features that are used to show support or to indicate interest in a post may be used in conjunction with one another.

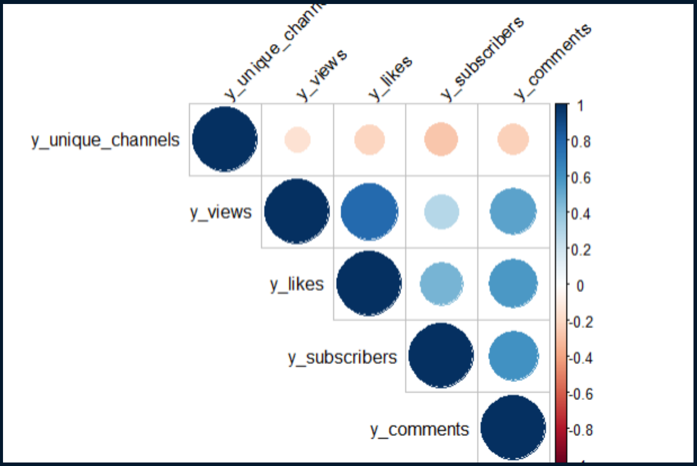
Reddit:



Twitter:



Youtube:

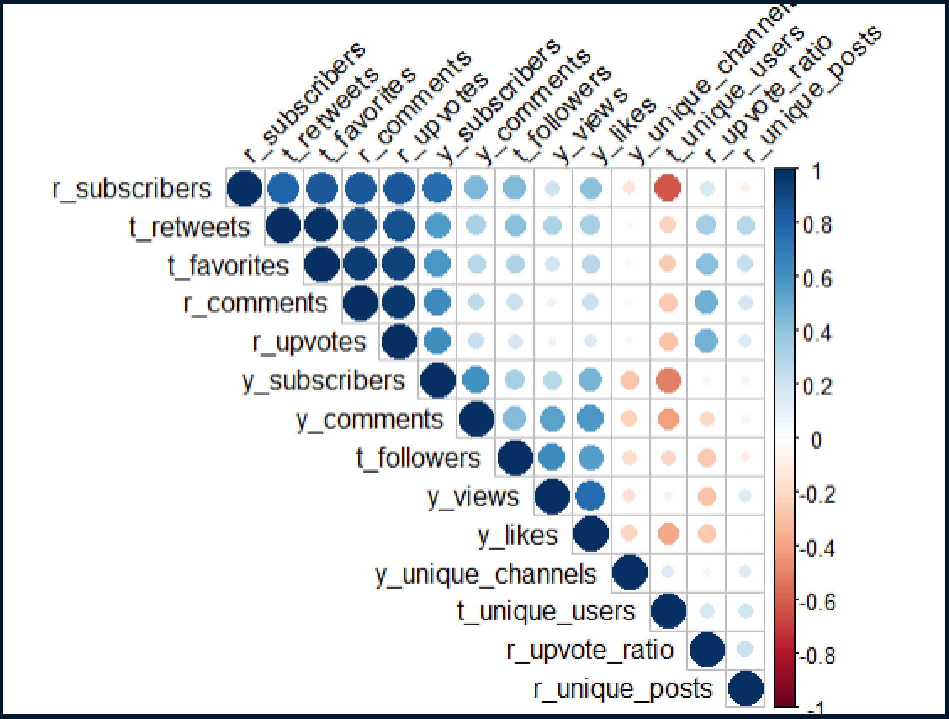


Reddit: we observed high levels of positive correlation between comments and upvotes. In other words, posts that generate a lot of thought and compel users to leave remarks will receive more upvotes. This also fits with the not as correlated relationship between upvote ratio and comments, as not all the attention on the post will be positive.

Twitter: Retweets and favorites are highly correlated in Twitter. Posts that resonate with more users also tend to be the ones that they share the most. It is also interesting to note that all variables are negatively correlated with unique users. This could be due to the fact that the more unique users who post in the top 10, the more variance there is in their followers, suggesting that individuals have a higher degree of influence on Twitter than other platforms.

Youtube: YouTube has low levels of correlation across most of its variables. Like twitter, the number of unique users within the top 10 videos is negatively correlated with all variables, most likely for the same reason. Views, likes, and comments are also positively correlated with each other as the more views something has, the more likely it is people will leave an impression.

In addition, we also examined cross-platform metric correlations to understand the interaction among social media in general.

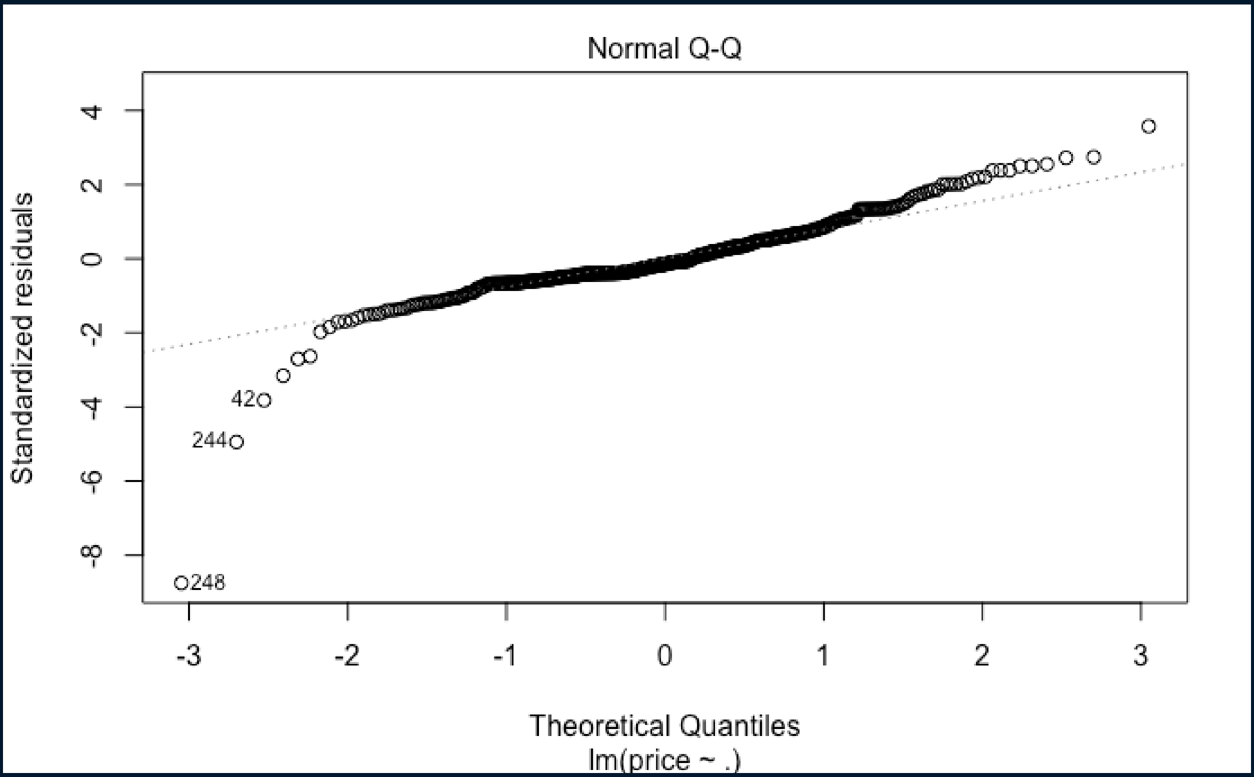


The most interesting feature is that Reddit and Twitter subscribers, retweets, favorites, and comments exhibit a high degree of correlation. This is likely due to the two platforms being more similar online news/blog information websites, as opposed to YouTube which is more of a video sharing platform. Most other factors aren't nearly as correlated, though Reddit Upvotes seem to be negatively correlated with unique twitter users within the top 10 posts. Observationally, the two do not exhibit any explainable correlation, so it is most likely due to variance from our smaller sample size (as will be noted in the limitations section).

Modeling

Modelling Method 1: Ordinary Least Squares and Shrinkage Methods

We began our analysis with an ordinary least squares regression on price using fifteen explanatory variables. We checked for normality using the Q-Q plot featured below. This gave us confidence in moving forward without any response variable transformations.



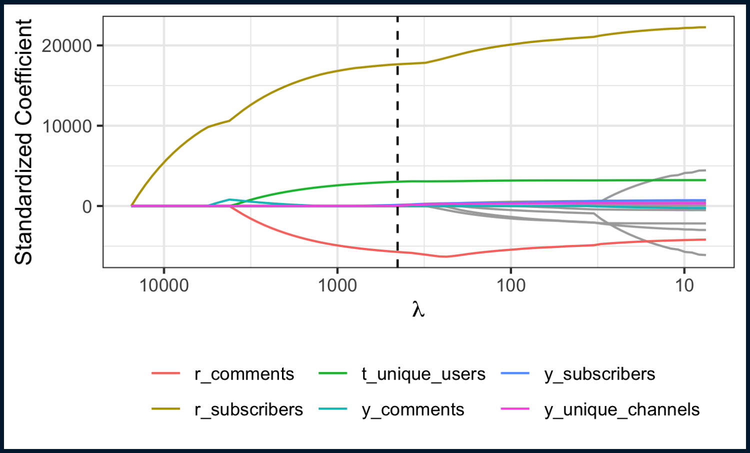
The variables that are significantly associated to the response at the 0.05 level are: r_subscribers, r_comments, r_upvotes, r_upvote_ratio, r_unique_posts, t_retweets, t_favorites, t_followers, t_unique_users, y_subscribers, and y_unique _channels. The multiple R-squared value indicates that these features explain 97.9% of the variation in response.

Ridge Regression and LASSO Regression

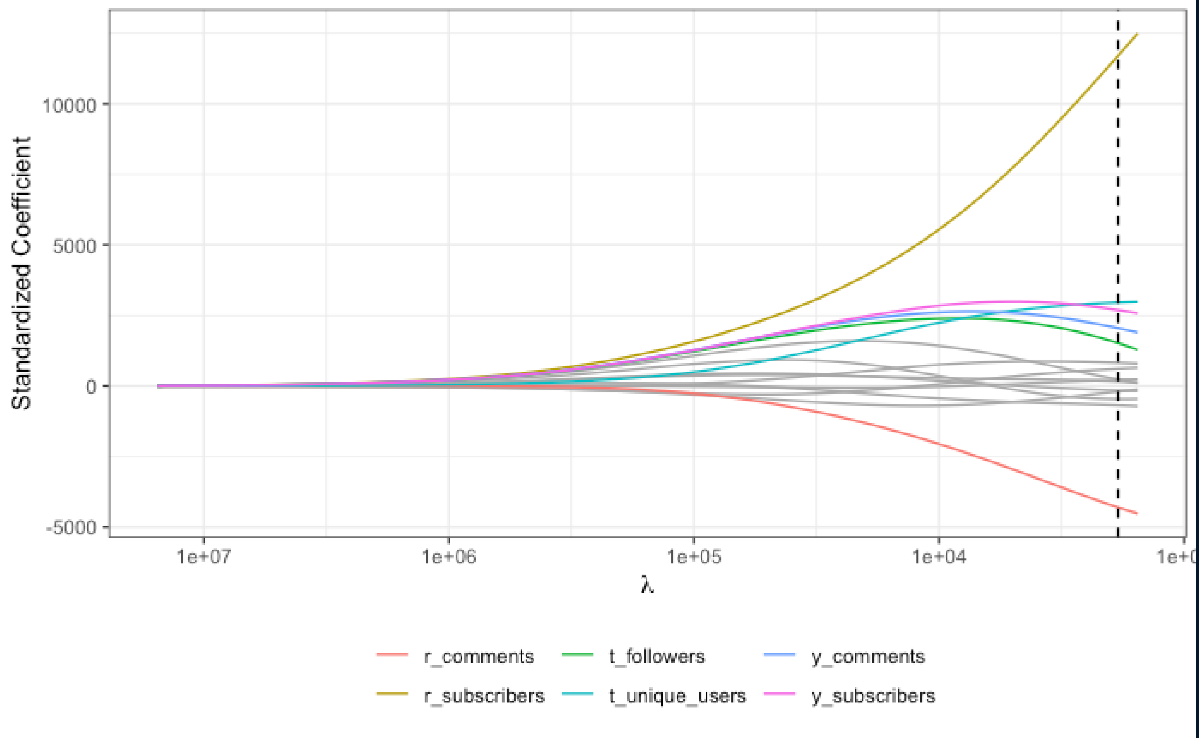
Our ordinary linear regression model produced a good model for the response variable. However, we wanted to try shrinkage methods on our model to see if we can improve the model and better determine which variables are more explanatory. The variables selected by LASSO are: r_comments, t_unique_users, y_subscribers, r_subscribers, y_comments, and y_unique_channels. Ridge regressions shrink all variables and does no selection of variables.

The plots for our LASSO and Ridge regression are featured below with the top six features highlighted:

LASSO:

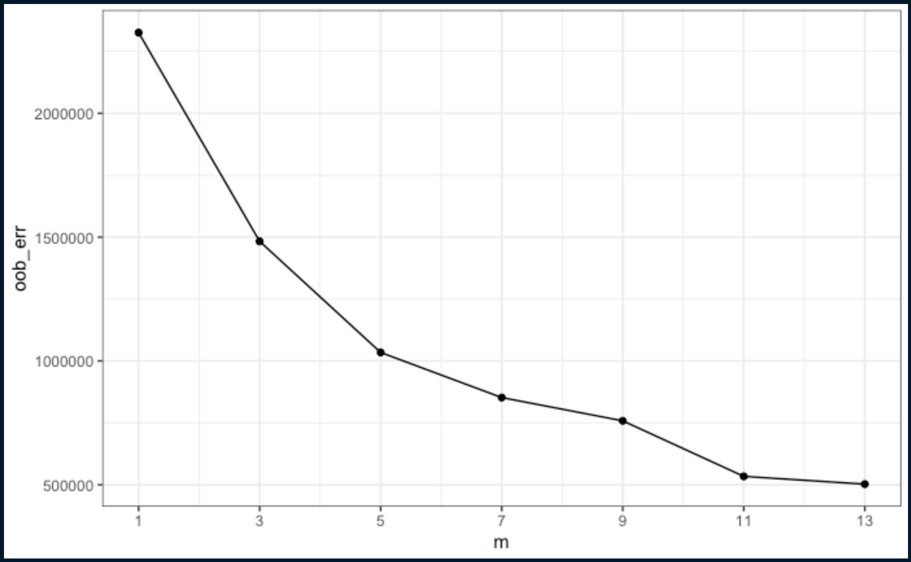


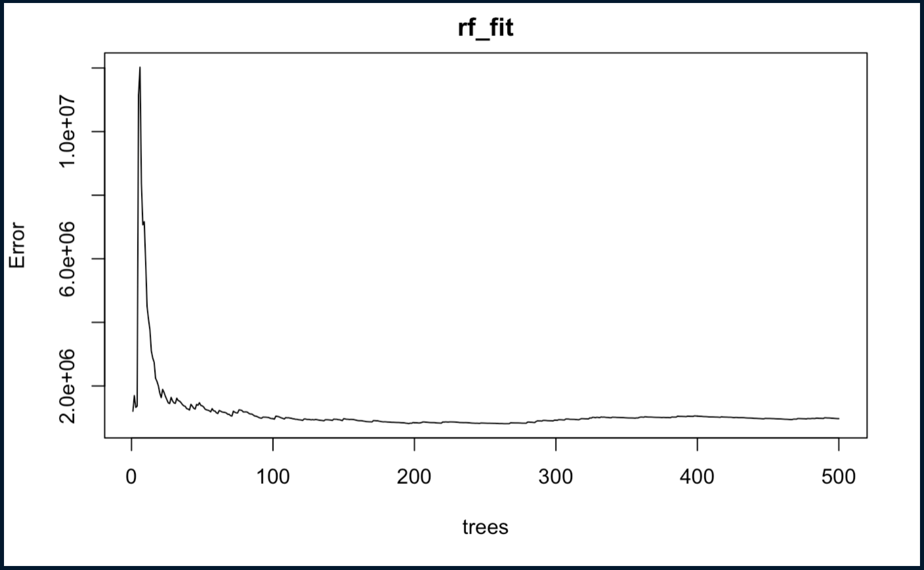
Ridge:



Modelling Method 2: Random Forests

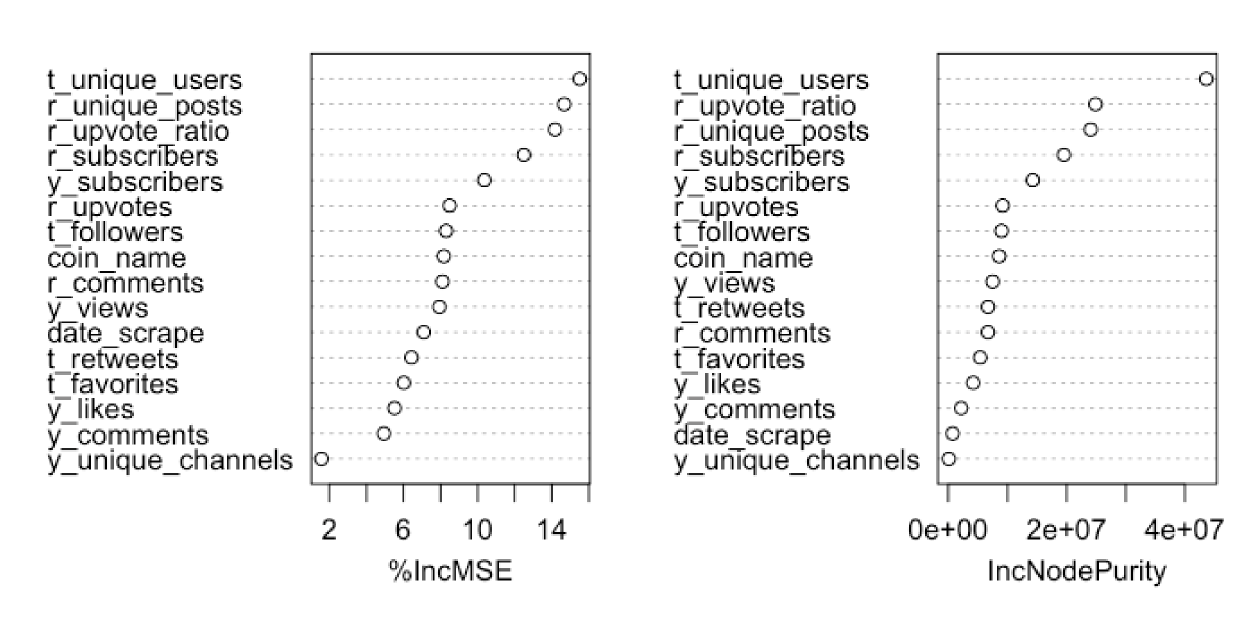
We then fit a random forest to the data. We decided to tune the random forest model for the optimal value of m, or the number of features to consider at each tree split, by trying various values of m and plotting the lowest out of bag error. This tuning was done with the intention to reduce variation. The out of bag error for each value of m can be observed below, we tested different values of m 1 to 13. Out of bag error was minimized at 13 with a general downward trend toward 13, giving us confidence in using 13 features for our random forest model.





After fitting our tree using the optimal m value at 13 as specified above, we assessed variable importance. For random forests, there are two notions of variable importance: purity based importance and OOB variable importance. OOB variable importance measures the decline in prediction accuracy after shuffling various features out of bag. Purity based importance measures the degree of improvement in node purity that results from splitting on given features. The results of both variable importance measures are summarized in the plots below.

Random Forest Importance



Based on the purity based importance and OOB Error improvement plots above, we saw that unique twitter users posting about a certain coin and average upvote ratio of top posts regarding a coin have the highest importance in determining price as measured by both metrics. There are also a number of reddit metrics that seem to provide high improvement. These are the most important variables in predicting coin price according to our random forest model.

The importance of each variable according to increase in node purity can be seen in the table below:

Feature	Increase to Node Purity
Date	58 M
Coin	81 M
Reddit Subscribers	63 B

Reddit Comments	78 M
Reddit Upvotes	41 M
Reddit Upvote Ratio	27 M
Reddit Unique Posts	8 M
Twitter Retweets	38 M
Twitter Favorites	40 M
Twitter Followers	65 B
Twitter Unique Users	117 M
Youtube Subscribers	22 M
Youtube Views	16 M
Youtube Comments	208 M
Youtube Likes	17 M
Youtube Unique Channels	5 M

Model Comparison

The Root Mean Squared Errors (RMSEs) of each of our models is presented below:

Ordinary Least Squares Regression RMSE: **2328**

Ridge Regression RMSE: 39294

LASSO Regression RMSE: 2635

Random Forest RMSE: 1607621220

Conclusions

Method Comparison

As shown from the RMSEs in the previous section, the ordinary least squares regression has the least test error with an RMSE of 2328, although all the test errors were alarmingly high. To address the first point, the ordinary least squares regression likely did the best because the true nature of the data is linear and additionally the low amount of features made LASSO and Ridge unnecessary. To address the second issue, we think our models were excellent at predicting the training data but performed poorly in predicting the test data. We'll explore why this might be the case in the *Takeaway* section.

For our Ridge and LASSO regression, we noticed the strongest variables were mostly overlapping. The overlap extended to the important variables chosen by the random forest. Overall, reddit metrics involving upvotes like average number of upvotes and upvote ratio seemed to be strong variables, and reddit metrics overall seemed to have large prediction value.

Takeaways

Our results suggest an aggregate of social media metrics is not a strong indicator of cryptocurrency prices. The lowest test error regression model, ordinary least squares regression, still produced a high RMSE. This suggests that social media metrics from a given day are not strong predictors of cryptocurrency price that same day.

Despite this, the overlapping metrics that exhibited strong predictive power on the training data suggest that certain metrics are relatively stronger than other metrics at predicting prices. Both LASSO and Ridge regression models suggested reddit comments and reddit subscribers are a strong predictor, and the random forest regression suggested reddit subscribers and reddit upvote metrics were strong predictors. This suggests that while social media metrics are not good predictors of cryptocurrency prices, reddit metrics may be relatively stronger predictors than other social network metrics.

Finally, while our high test errors suggest that social media activity does not have a large correlation with cryptocurrency prices, it could be the case that alternative approaches would have stronger results. For example, if we were to focus our data on individual coins, the social media metrics may serve as stronger indicators. Another example, if we were to conduct transformations on the data putting all the price data on the same scale or shifting the price data down to see if current social media activity can predict future activity, we may have seen better predictive power and lower test errors. We will explore these potential methodology changes in the Future Steps section.

Overall, cryptocurrencies continue to be a growing market industry, and strategies that use social media data to trade are becoming increasingly more commonplace. Our analysis suggests that social media based strategies to trade can't use current social media activity to predict current prices across the whole market. We recommend strategies that contextualize based on coins or use current social media data to predict future prices.

Limitations

Data Limitations

The main limitation of our data was that the sample size is quite small, especially when considering the implications of our exploration. Two months of data, in the scheme of the decade that some of these coins have been in existence or even the three years since their resurgence, is a tiny portion of their data. A small sample size produces a high degree of variance, meaning if we sample another 2 month period, the results could be drastically different as we aren't capturing as much of the underlying trend within the coin prices. There is some discussion about how far data collection should go, as market conditions and some structural changes in the way crypto currencies have been deployed have made market conditions in the past three years different than how they were in the seven years prior. As such, more care needs to be taken when choosing an adequate observation period. Additionally, cryptocurrencies are a global currency, and may be traded in different places. Not only were we not able to include every social media platform, even within the US, we were also unable to capture data from country-specific social media platforms. In smaller countries where the crypto market is actually larger than their national currency, understanding their sentiment towards crypto may be important as they become reliant on crypto in the future.

Analysis Limitations

From an analysis perspective, our models aren't directional, i.e., they can't gauge whether social media attention around something is positive or negative. For example, Elon Musk's negative tweet about Bitcoin generated a lot of attention on twitter, increasing the number of retweets and favorites. Instead of classifying tweets based on their shown sentiment, our model averages between positive and negative tweets, and spits out a middle number that attempts to capture the in-between effect of those social media metrics, which isn't necessarily useful when predicting coin prices in particular circumstances. However, Elon Musk isn't representative of the entire crypto following on social media, and our data accounts for the overwhelming majority of individuals who make up the crypto community on social media. This means that there is still some explanatory value in our predictors, though it is important to address the actual content of the social media posts themselves.

Additionally, our analysis was based on the top posts of each crypto currency, but didn't factor in a lot of the smaller ideas and thoughts being shared around the internet. While headliner items are intuitively the most impactful on crypto prices, smaller opinions and micro-communities make up a non insignificant portion of thought regarding crypto. Because crypto began as a niche community, the remnants of these early communities may include investors with large holdings of these currencies, and therefore have a disproportionately higher effect on price.

Future Steps

Improving the quality of the dataset is the most immediate and actionable step in improving the quality of our model. We could expand the two month period to three years in order to account for the recent boom in cryptocurrency. This would allow us to tune out the seven years prior, which included tremendous peaks and troughs that would've lowered the accuracy of our model based on current market conditions. Additionally, we could include more social media platforms into our consideration. Notably, Facebook and WhatsApp are left out from our list. Small coins have recently begun promotion campaigns via Instagram and TikTok for their coins, so accounting for new marketing methods would also help in understanding changes in coin prices. Also, though China has banned cryptocurrency, other countries around the world are adopting crypto, and each have their own set of social media platforms that we have not accounted for.

It is also important to address the nature of social media posts, rather than just accounting for exposure. In order to account for the content of the social media posts, we could implement a Markov chain, or similar machine learning algorithm, that is able to track the sentiment of the social media post based on word associations (for Twitter and Reddit), and audio translation (for YouTube). This would allow us to create a classifier for "positive", "negative", and "neutral" attention, which would allow the model to be tuned for particular types of social media attention. This would make it more predictive when a particular scenario arises, rather than just taking the average of all three sentiments on social media.