The Effects of Google Search Trends on Cryptocurrencies

Final Report

Problem Statement

Since Cryptocurrencies first came onto the market in 2009, they have garnered an immense amount of praise and criticism. Supporters of cryptocurrency have heralded them as being a decentralized currency that has protection from inflation and is self-governed and managed. Critics have expressed concern over its environmental impact, its uses in illegal activities and its high volatility.

Whichever your stance on cryptocurrency may be, it is certainly a topic that many people have expressed their opinion on online. In fact, recently Elon Musk has made headlines by announcing that Bitcoin, Ethereum, and Dogecoin was going to be an acceptable mode of payment for their future project of landing on Mars. Having all this attention and people talking about these coins shot the prices up overnight to record highs.

I think that this is an aspect of cryptocurrencies that is worth exploring. Do people's interest in cryptocurrencies affect the price of the coin? And can we use this data to make predictions about the price of cryptocurrencies?

Datasets

- CCi30 Index – The CCi30 is a rules-based index designed to objectively measure the overall growth, daily and long-term movement of the blockchain sector. We were looking to mimic this index
- Top 30 cryptocurrencies from Yahoo Finance – To mimic the CCi30 index we pulled data from the Yahoo Finance API. To measure the price of stock, we pulled daily 'Closing' price of each stock.
    - Stocks included in our data: ['BTC-USD', 'ETH-USD', 'USDT-USD', 'BNB-USD', 'ADA-USD', 'DOGE-USD', 'XRP-USD', 'USDC-USD', 'DOT1-USD', 'HEX-USD', 'UNI3-USD', 'BCH-USD', 'LTC-USD', 'SOL1-USD', 'LINK-USD', 'MATIC-USD', 'THETA-USD', 'XLM-USD', 'VET-USD', 'ETC-USD', 'TRX-USD', 'ICP1-USD', 'FIL-USD', 'XMR-USD', 'EOS-USD', 'AMP1-USD', 'AAVE-USD', 'SHIB-USD', 'ALGO-USD', 'CRO-USD']
- Google Search Trends Dataset – Looked at popularity of search terms from top crypto currencies- "Cryptocurrency", "Bitcoin", "Ethereum", 'XRP", "Dogecoin".

Data Wrangling

Originally, I was pulling daily stock data from Yahoo Finance API from 1/1/2015 to 8/1/2021 which gave us 2401 rows and 30 columns. However, there were many missing values present. Many cryptocurrencies did not exist in 2015. Some columns of stock closing prices that had over 1500 or more missing values. Removing the rows and columns with many missing values left us with 1025 rows and 19 columns. For any other missing values found in the middle of the dataset, we used forward fill to take the last closing price and fill missing values with those values.

Our Google Trends Dataset measured the popularity of different google search terms measured as weekly data. To match my stock price data and trends data, I resampled Yahoo Stock Data from daily measurements to weekly measurements

EDA

The first step in our project was to mimic the CCi30 index. The CCi30 index takes in data from the top 30 crypto currencies and returns a value that represents the overall health of the cryptocurrency sector. It is formatted like a typical historical stock data table with “Date”, “Open”, “Close”, ”High”, ”Low”, and “Volume” as columns. Here we only use the closing price of the stock as a metric to evaluating the valuation of the stock.

The closing price is typically used in stock market valuations as the most accurate measure of stock performance. Seen in Figure 1.

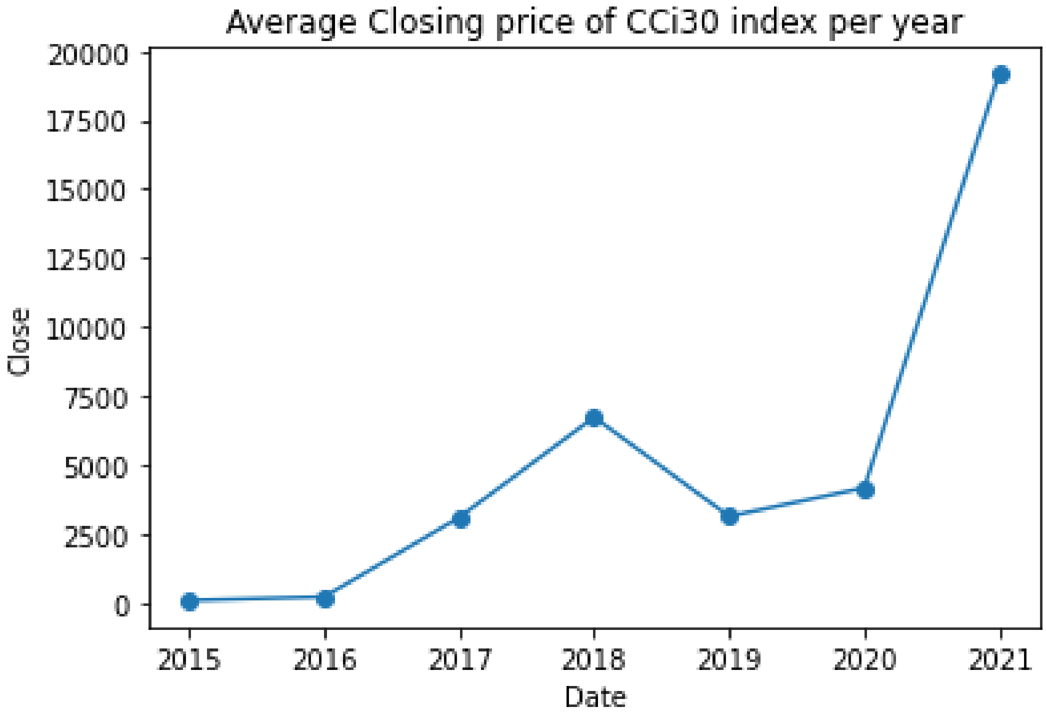

Figure 1: Closing Price CCi30

Google search trends all follow a similar structure. It is no surprise that specific terms such as bitcoin has the most searches. It is the largest and most popular cryptocurrency on the market.

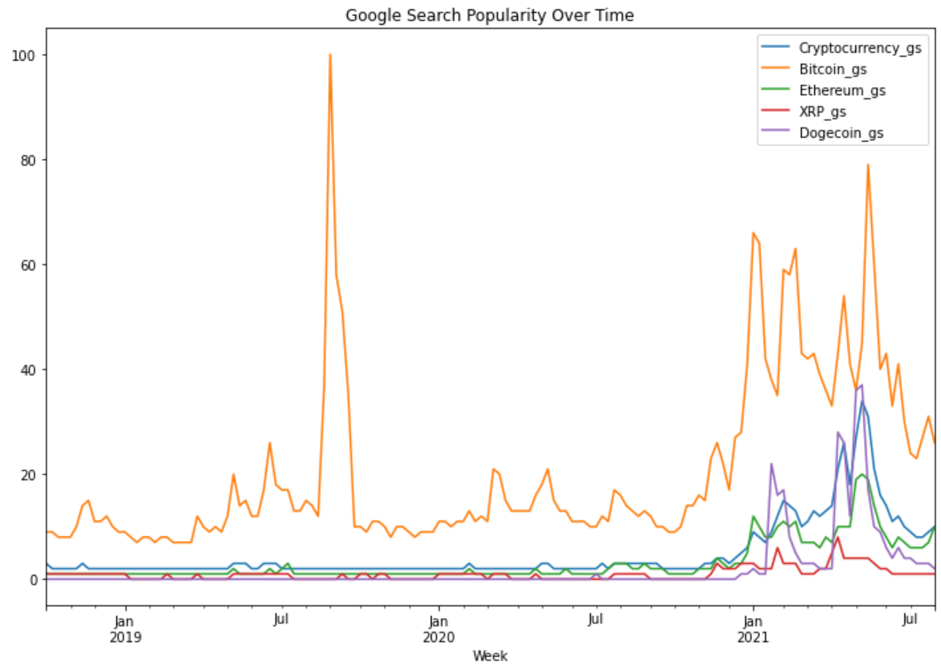

Figure 2: Google Search Trends

PCA Analysis

I ran a PCA analysis on Yahoo Stock Prices to reduce the number of features from 30 down to 4 and decorrelate the features. Our PCA analysis showed an optimal number of 4 components that explains 92% variance of the distribution.

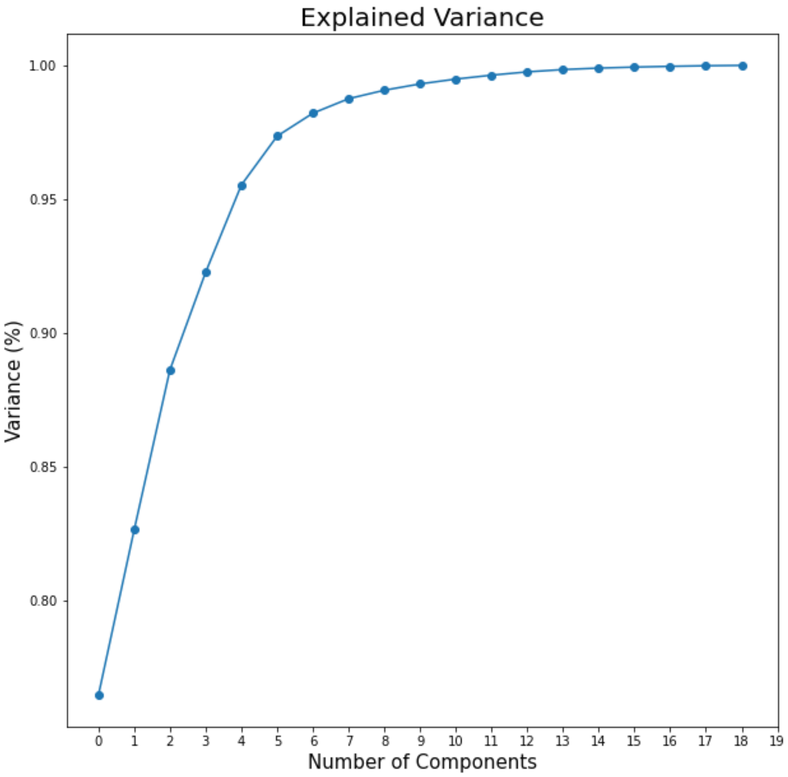

Figure 3: Scatter plot showing the percent variance explained by number of components

Using our PCA analysis we now have the eigenvectors of our 4 principal components for each week of our sample. Eigenvectors explain which direction (either positive or negative) our stock prices will go in our "CCi30 index". To use PCA as a visualization tool, we took the row wise mean of our 4 PCA components and plotted it in a graph

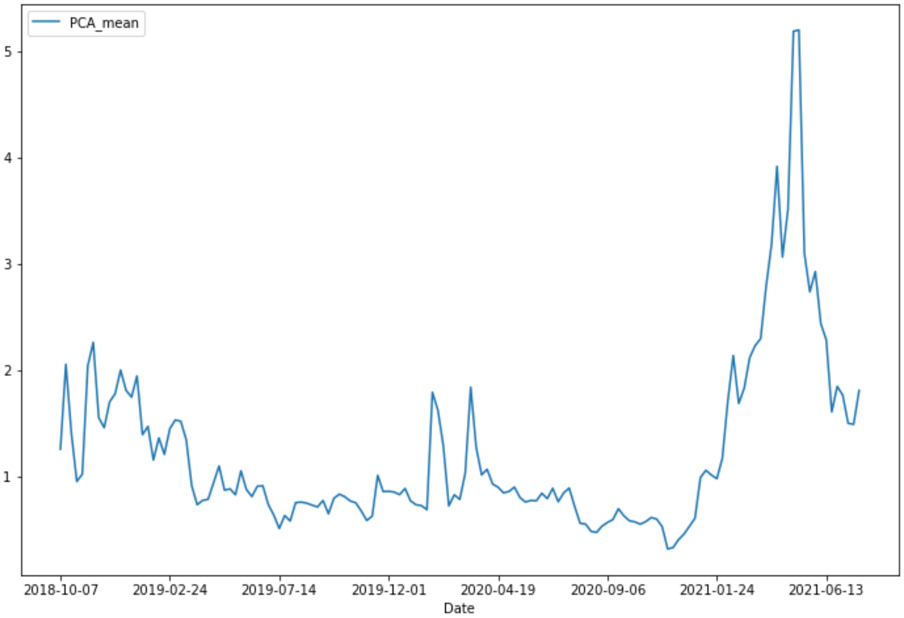

Figure 4: Visualization of our PCA Analysis

We can now use this data as a representation of closing prices for cryptocurrencies from 2018 – 2021. We will be using these values in our PCA analysis to compare them with the popularity of specific google search terms.

Model Selection

As a precaution, I checked the correlational matrix of my independent variables (google search terms) for multicollinearity and found that while all the features are correlated with each other in some way, there were some features that were too highly correlated with each other and would affect the results of our model. We used a threshold of .80 and above as our measurement for collinearity.

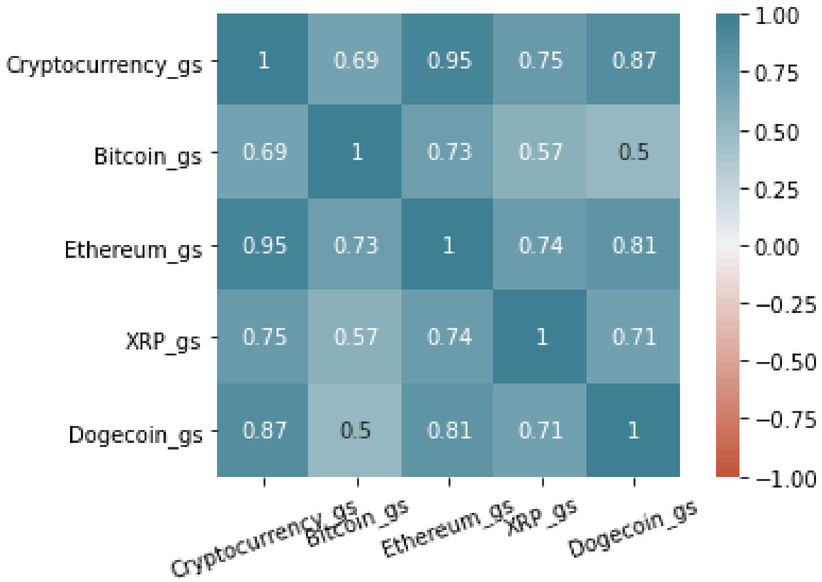

Figure 5:Correlational Matrix of my Independent Variables

To solve my issue with collinearity, I can choose to remove an independent variable and keep the other, or I can create a new column with the average of both the variables and drop both correlated variables. In this problem, I dropped the cryptocurrency google search column as it was highly correlated with multiple variables and took the average of Ethereum and Dogecoin columns.

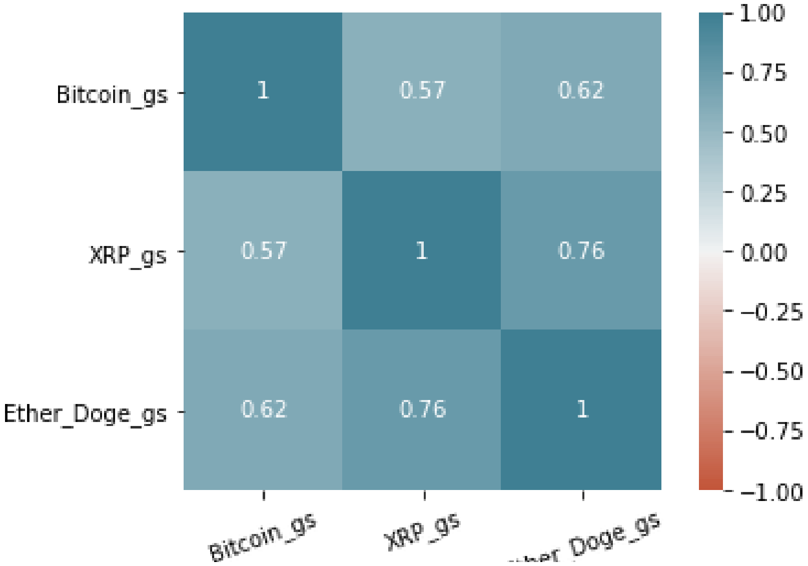

Figure 6: Correlational map with collinearity fixed

For the model we tried three different regression classifiers: Linear Regression, Random Forest Regression, and Ridge Regression. We are looking for a model that can predict on our test data and has the highest r squared score with low MSE.

Note that my data is a time series, so my cross validation must be done differently. Normal cross validation splits the train and test set randomly. You may have data that has occurred in the future in the training set, used to predict on past dates in the test set. To fix this problem, I implemented a rolling training set and test set for my data, to properly consider the time series dependency of my data.

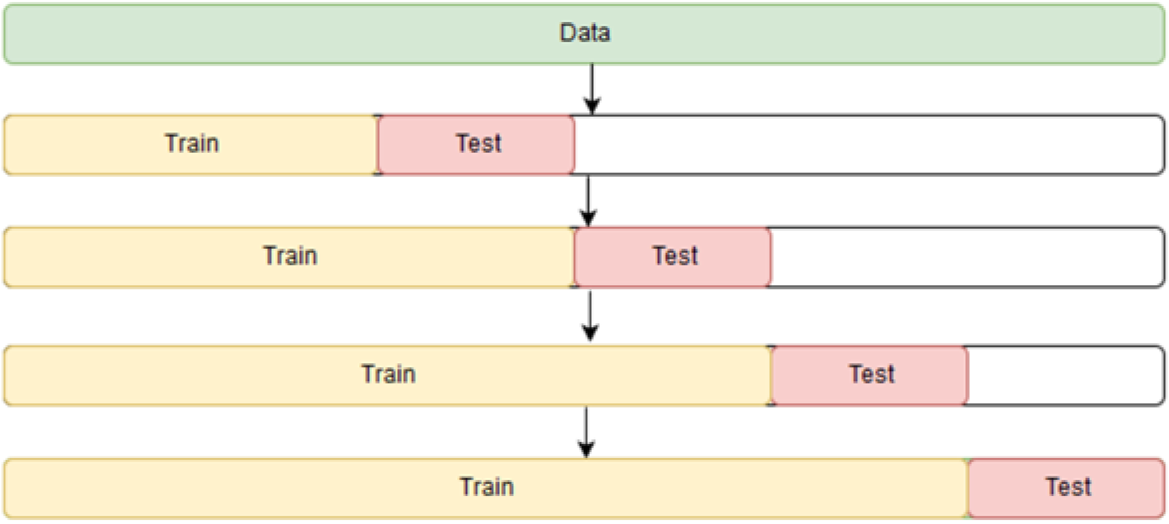

Figure 7:Cross Validation for time series data. (source: https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4)

I used the sklearn package TimeSeriesSplit to split my data 3 times and ran each model on the training set and ran a prediction on the test set. I take each r squared and MSE score from the splits and average them to find the overall r squared and MSE of each model.

Linear Regression Results

R squared was **0.198** for our cross-validation splits with a mean MSE of the training data of **0.142** and a mean of the test MSE of **19.32**. This is a poor model to use for this dataset due to a low score in r squared and a large number in our mean test MSE. That means that the sum of errors of our predictions is approximately 19.32 from the line of regression.

Ridge Regression Results

Ridge regression performed much better. R squared was lower in this model with a score of **0.11**. However, the mean MSE of the training set is a **0.16** and the mean MSE of the test set of **2.93**. I still would not choose this model, because while the lower MSE score means the predictions were more accurate, the lower r squared score means that there was a problem with how much variance our independent variables explain the dependent variable.

Random Forest Regression Results

R squared was much higher in this model with an average score of **0.46** for our cross validation splits with a mean MSE of the training data of **0.095** and a mean of the test MSE of **1.19**. This is our best performing model and the model I ended up choosing at the end. It has the highest r squared score and the lowest MSE of both the train and test set.

I also plotted the feature importance of our model. Bitcoin google searches has the biggest impact out of all the independent variables.

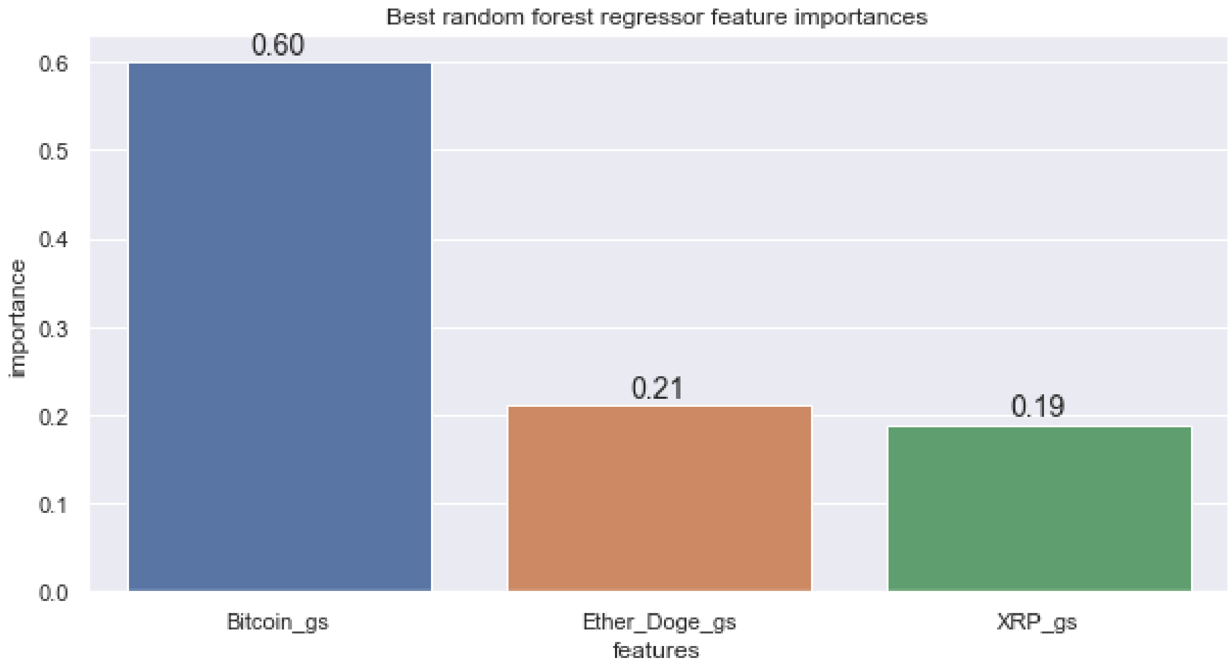

Figure 8: Feature Importance from Random Forest Regression

Recommendations

- The model shows that google search trends does have some impact on the price of cryptocurrencies. Keep people's interests high on this topic may prove to be healthy for the cryptocurrency market
- Bitcoin's affect on the market is significant. Bitcoin dominates in both stock price and google search trends. Expect the price of the coin to increase when more people are talking about it. As Bitcoin starts to reach its maximum capacity, expect more conversation to be generated over the remaining limited supply.
- Bitcoins feature importance is a problem in this model. Its weight is too high to see the effects of the other features on the data. The overall health of cryptocurrencies is highly dependent on how bitcoin is doing. This skews the data to lean towards bitcoin for our predictions

Further Research

Further research can be done on expanding the possible variables that measure people's interest in crypto currencies. Accessing tweets through the twitter api and looking at hashtags could provide some more insight on the type of discussion that users are generating. This can be further investigated to classify if discussions are positive or negative by using NLP algorithms.

We can also categorize people's interest via age and gender. It would be interesting to which groups of people are talking about cryptocurrencies.

There has been an increasing amount of “scam coins” entering the market recently. These are coins that are heavily promoted via social media, luring people to invest heavily in the coin, inflating the price, only to have the owners sell all their stock and leave with the money. What if the model can measure the amount of interest a coin is generating and determine if a crypto currency is legitimate or a scam?