Michael Rakhamimov and Jonathan Ha
INST737
12/10/2022

Crypto and Stock Predictive Analysis Document

Introduction:

In a world full of uncertainties, people gravitate towards sources of information in order to make knowledgeable decisions. This concept is especially prevalent in the field of investing. When an individual chooses to invest in a company or a token, it is advised that they do as much research as they can before staking their money. With that in mind, we chose to create predictive models of crypto and stock data to help investors make more informed decisions when investing.

When investing in fungible tokens, there is the traditional method of investing in the stock market and then there is the more modern alternative of investing in crypto currency. Since cryptocurrency is heavily dependent on technology, the most relevant stock that we can draw comparisons from is NASDAQ as it is responsible for most major publicly trading tech companies. In order to represent the entirety of the NASDAQ umbrella, we chose to use the NASDAQ composite index as the stock variable. For crypto we chose 5 unique, popular crypto currencies: Bitcoin (BTC), Ethereum (ETH), Ripple (XRP), Litecoin (LTC), and Binance (BNB).

We collected the raw data on the 5 cryptocurrencies that we identified above as well as the NASDAQ composite index. After cleaning and merging the datasets, we conducted two separate analyses, one in R and the other in Python. The cleaned data was imported to R where we ran several simple and multiple linear regressions to test the relationships of these assets and ultimately identify any correlation among them. Additionally, we created a new variable in R called Quarters, which defines the fiscal quarter for each entry so we could identify any seasonality among the data points. In Python we performed the necessary steps to complete the granger causality test on our data set, to further support the correlation claims we found in the R analysis. These steps included running ADF and KPSS tests on the data to identify if the time series is stationary or not.

Regarding specific contributions from each member, the overall work was shared evenly. Michael Rakhamimov took care of the initial data collection, cleaning, and merging required to transform the raw data into data that is ready for analysis in R. Additionally, he wrote the code for the analysis within R. Jonathan Ha also conducted some data cleaning and integration in order to get the dataset to work properly in Python. In addition, he wrote the code for the analysis that was done in Python. Work on presentation material was shared evenly as well.

Literature Review

Linear regression is important in financial analysis as it is useful in seeing if the value of various assets correlates with one another, as well as if one asset may be a good predictor in the price of another asset. In a study similar to ours, researchers looked at how Bitcoin may correlate with the stock market and whether it is a good predictor of changes in the stock market. Within their study they created a predictive model with Bitcoin as the predictor and tested it using historical data concerning stock returns. Their results indicated that Bitcoin was a good predictor of stock market returns [1]. This insight was helpful as it supported both our expectations and results, as we felt that due to Bitcoin being a well-developed and popular cryptocurrency, it would correlate strongly with the stock market. Their conclusion however was a bit too generalized as they believed that digital assets as a whole could be used to predict stock returns. We felt as though this was a bit too broad of a conclusion to draw from their

study which focused on only one cryptocurrency. This is because there are thousands of cryptocurrencies that can highly vary from one another, especially if they are not as developed as more popular ones such as Bitcoin.

In addition to linear regression, Granger causality analysis is also used widely in the financial field as it helps to show causality between various financial markets as well as specific assets such as if certain stocks may affect one another. Although it does not indicate 'true' causality, it has still been shown to be beneficial in showing if the change in value of one asset may affect the change in value of another asset. This is important as it allows for researchers to discover whether a pair of variables may "cause" one another. Use of Granger causality may be beneficial in many cases, including but not limited to stock market analysis, research towards what variables affect developing economies, and discovering market risks.

There have been previous studies where the Granger causality test has been used to look at the relationship between cryptocurrency and the stock market. In one study the relationship between cryptocurrency and emerging stock markets was looked at using the Granger causality test as well as the Liang causality analysis. The study used the CRIX index, which is an index created from the top 30 cryptocurrencies as well as the MSCI Emerging Market index. From their use of the Granger and Liang analyses, it was determined that the cryptocurrency market and emerging markets start off independent of one another, however, as they grow, they gradually become more connected with one another. Additionally, they once again are shown to become independent of one another once they become stable after an intense period of fluctuation [2]. Although the study focused on emerging markets, whereas our study focuses on a developed market, it still provided many insights for us and helped to support our results. This is shown by the fact that as the emerging markets and cryptocurrency grew, they became more closely related to one another, and since a developed market has had the time to grow already, it would make sense for it to have a relationship with cryptocurrency.

Methodology

As previously mentioned, we collected the time series data of 5 cryptocurrencies and the NASDAQ composite index. The datasets for the cryptocurrencies were individually downloaded as CSV files from a popular cryptocurrency data aggregation platform CoinGecko [3][4][5][6][7]. The dataset for the NASDAQ composite index was downloaded directly from NASDAQ's website as a CSV as well [8].

Once all the csv files were downloaded locally, they were compiled into one data set in Excel and merged using a join statement in Power Query on the corresponding date of each entry. In this process, the data is also cleaned for any missing date entries as they are filtered out in the merge.

Because these assets can be bought with a variety of different currencies it is important to note that the prices that we mention for crypto the cryptos as well as the price of the NASDAQ composite index are both in USD. Although the variable names do not contain this annotation for the sake of concise nomenclature, the labels on the plots generated indicate the value is represented in USD.

Our solution to uncertainty when investing was the creation of several predictive models using a combination of simple and multiple linear regressions. In the R code we defined the models for all the variable combinations in the simple linear regressions. For the multiple linear regressions, we defined multiple models with 2 explanatory variables as well as one model with 3 explanatory variables. The models with a strong goodness of fit could be used to assist investors when choosing a future asset to buy.

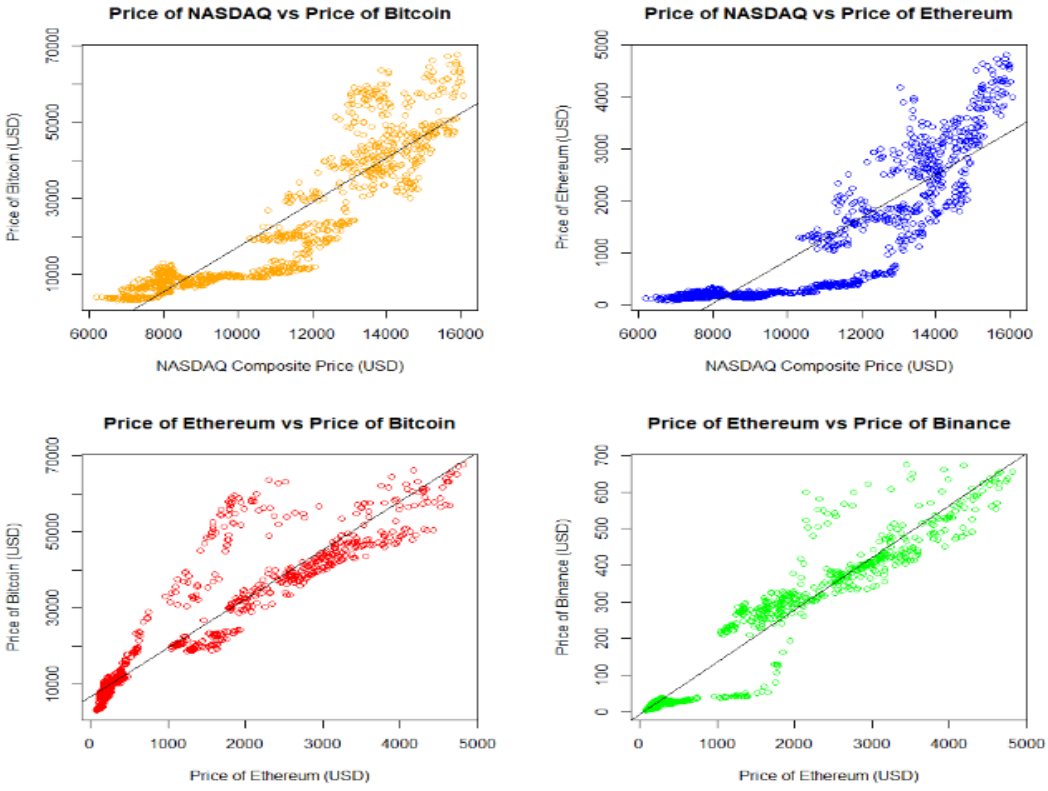
Although these models show a definitive correlation for several variables, this still is not enough to establish causality. With that in mind, we conducted a separate analysis in Python where we used the

granger causality test to support our claims from the R analysis. All the data and code is then backed up into a repository [9], which will be submitted along with the rest of the deliverables.

Experiment Result

In order to gain insight on the seasonality of the data we created the custom time variable Quarters in R which identifies the fiscal quarter that entry belongs to. With all these variables loaded into an R data frame, we intended to make predictive models using simple and multiple linear regressions. The relationships between the explanatory variables and the dependent variable are evaluated by the results of the linear regression. If all the explanatory variables including the intercept have a p value that is less than our alpha of .001, the relationship is considered significant. From there we look at the R squared value to identify the goodness of fit. A moderate to strong goodness of fit can be interpreted from any R squared value that is larger than .75. If the variables are all significant, the p value of the model is significant, and the goodness of fit is moderate to strong then the model shows the variables are correlated.

Results from our analysis in R yielded several significant relationships but only a select few had a strong enough goodness of fit to be considered for actual use. 4 noteworthy relationships worth highlighting are the relationships between Ethereum and Bitcoin, NASDAQ and Bitcoin, NASDAQ and Ethereum, and finally Ethereum and Binance. Each of these models yielded significant p-values for the variables and model as well as a strong goodness of fit. The scatter plots below depict these strong positive relationships, and the line of best fit is representative of the linear regression equation that belongs to each model.



Our findings from the R analysis indicate that there is a correlation between the price of cryptos and the price of the NASDAQ. Additionally, we found a correlation among several cryptocurrency relationships with other crypto. The noteworthy relationships we found have been visualized above.

However, it is important to note that the regression models can only yield results about correlation and that further research was required to define causality.

We had originally planned on focusing on just correlation, however with suggestions from our professor, we also ended up performing the Granger causality test which looks into "causation".
As we did not have knowledge on how to perform the Granger causality test, we looked up guides on how to perform the test. After looking at various resources, we discovered that prior to performing the test, it is important to see if the dataset is stationary or not. The process behind this is performing includes performing the ADF-test and/or the KPSS-test. Both tests test for stationarity although their null and alternative hypotheses are opposite one another. The ADF-test 's null hypothesis is that a time series is non-stationary, and its alternative hypothesis is that it is stationary, while the KPSS's test's null hypothesis is that a time series is stationary, and its alternative hypothesis is that it is that a time series is non-stationary. After performing the test, if the dataset is stationary, we have to perform the difference method which would make it stationary. Stationarity is important as it helps to get rid of trend and seasonality, which is helpful as it makes observing datasets easier. After ensuring that the dataset is stationary, we can then actually perform the Granger causality test. When performing the test however, the number of lags must be set, which is basically the number of past x values looked at and how helpful they are in predicting the response values, y.

When actually performing the ADF- and KPSS- tests on our datasets, we found that for the ADF-test, all variables other than XRP were stationary, whereas for the KPSS-test, all values including XRP were stationary. Although there was this contradictory result, we chalked it up to the limitations of the tests, which include the dataset being possibly too small. As the data was non-stationary, we performed the difference method which made it stationary. After performing the test, we re-performed the test, and all of our variables did indeed become stationary. With our data stationary, we were able to perform the Granger causality test. With a lag of 15, we noticed that the majority of variables 'Granger caused' caused one another. The only exception to this was that the price of BNB, XRP, and LTC did not 'Granger cause' the NASDAQ average. This was within our expectations however, as BTC and ETH, both of which were shown to 'Granger cause' changes in the NASDAQ average, have been the most popular cryptocurrencies and have had the most time to grow.

		Predictor					
		BNB_price_x	XRP_price_x	LTC_price_x	ETH_price_x	BTC_price_x	NASDAQ_Average_x
Response	BNB_price_y	1	0.0106	0	0.0016	0	0
	XRP_price_y	0.0005	1	0	0	0	0.0003
	LTC_price_y	0.0001	0.0014	1	0.0001	0.0001	0
	ETH_price_y	0	0.0465	0	1	0	0
	BTC_price_y	0	0.0012	0	0	1	0
	NASDAQ_Average_y	0.1101	0.5964	0.449	0.009	0.0009	1

Limitations:

There were various limitations we faced throughout the process of our project. A couple of limitations revolved around our datasets. One limitation was that our datasets were limited by both date and type of cryptocurrencies. Our NASDAQ dataset was limited to the years 2012 through 2022, and our cryptocurrency dataset was limited to the years 2017 through 2022, as well as being limited to the cryptocurrencies BNB, XRP, LTC, and BTC. Another limitation was that some of the records within our cryptocurrency dataset and our NASDAQ dataset had different date ranges. Due to this, after combining the two datasets together to work with, we had to drop NA values from the merged dataset which decreased our total number of records. Next, although the Granger causality test tries to show causation between two variables, there are several limitations to the test. The limitations of the test include not being an indicator of true causality, being limited to the analyzation of two variables at a time, and there being misleading results when a variable's real relationship involves multiple other variables. Due to this, although our results showed that there was causality between various pairs of variables in our project, we cannot state that there is true causality. Additionally, although the predictive models we highlighted may be significant and have a strong goodness of fit, the models are sensitive to change with every new entry. As a result, future data will yield different results if plugged into the same predictive model constructed in R.

Conclusion and Future Work

Although true causality requires more research to establish, the analyses conducted in R and Python provide promising evidence that our work can help future investors. In R we were able to identify strong correlations between the crypto, stock, and quarter variables using various combinations of simple and multiple linear regressions when creating the predictive models. With a given accuracy defined by the R squared value, these models can be used to estimate the value of a given asset assuming the relationship is significant. In Python we prepared the data for the granger causality test by performing the ADF and KPSS tests to identify whether the time series is stationary. Following that we performed the granger causality test across several different lags to simulate any variance. Given all the correlations identified in the predictive models, in conjunction with the results from the granger causality test, an investor would be well informed to make the decision on which asset to invest in.

Regarding future work, we would like to expand upon the custom time variable that we created in the R code. In this instance we chose to only label the temporal data as the fiscal quarters that they fit into between the numbers 1 and 4. If we were to expand this quarter variable, we would make each new year the next 4 subsequent numbers following the quarters in the previous year. This would allow us to test for future outcomes when used as an explanatory variable.

References

[1] Lu, X., Liu, K., Liang, X. S., & Zhang, Z. (2020). The Break Point-Dependent Causality between the Cryptocurrency and Emerging Stock Markets. *Economic Computation and Economic Cybernetics Studies and Research. 54. 203.* 10.24818/18423264/54.4.20.13

[2] Isah, K. O., & Raheem I. D. (2019). The hidden predictive power of cryptocurrencies and QE: Evidence from US stock market. *Physica A: Statistical Mechanics and its Applications.* https://doi.org/10.1016/j.physa.2019.04.268

[3] CoinGecko. 2022. Bitcoin USD historical data. (December 2022). Retrieved December 10, 2022 from https://www.coingecko.com/en/coins/bitcoin/historical_data#panel

[4] CoinGecko. 2022. Binance USD historical data. (December 2022). Retrieved December 10, 2022 from https://www.coingecko.com/en/coins/bnb/historical_data#panel

[5] CoinGecko. 2022. Ethereum USD historical data. (December 2022). Retrieved December 10, 2022 from https://www.coingecko.com/en/coins/ethereum/historical_data#panel

[6] CoinGecko. 2022. Litecoin USD historical data. (December 2022). Retrieved December 10, 2022 from https://www.coingecko.com/en/coins/litecoin/historical_data#panel

[7] CoinGecko. 2022. Ripple USD historical data. (December 2022). Retrieved December 10, 2022 from https://www.coingecko.com/en/coins/xrp/historical_data#panel

[8] NASDAQ. 2022. NASDAQ Historical Data. (December 2022). Retrieved December 10, 2022 from https://www.nasdaq.com/market-activity/index/comp/historical

[9] Michael Rakhamimov and Jonathan Ha. 2022. Project Repository. (December 2022). Retrieved December 10, 2022 from https://github.com/Mishanya11/INST737_Group_Project