**Cryptocurrency Market Analysis**

**OPIM 5512 Summer 2019**

**Team 2**

**Team Members:**

Apurwa Prasad

Anthony Renzullo

Kumari Juhi

Xinyin Miao

Introduction

A cryptocurrency is a digital means of exchange created and used by private individuals or groups. Strong cryptography is used to secure financial transactions, control the creation of additional units, and verify the transfer of assets. Developers build protocols and complex code systems using advanced mathematics and computer engineering principles to make them virtually impossible to duplicate or counterfeit. Cryptocurrencies are also marked by decentralized control. Their supply value is controlled by user activity and complex protocols in their governing codes as opposed to conscious decisions of central banking systems. They are not regulated by national governments so they’re considered alternative currencies, that is, ways of financial exchange that exist outside the bounds of state monetary policy. Hundreds of cryptocurrencies exist today, but many are created each month (1). The top cryptocurrency, bitcoin, was first released in 2009. Other prominent coins are Ethereum, Litecoin, and XRP (2).

Cryptocurrencies might be best known for their volatility and uncertainty. Such fluctuations can scare away investors, especially companies. In January 2018, cryptocurrencies reached a market cap of over $795B with almost $300B of that from bitcoin (CoinMarketCap).

Literature

Though cryptocurrencies are a relatively new form of currency, there has been extensive research on the topic. Particularly, we focused on studies that involved predicting prices of these currencies. One such study out of Southern Methodist University analyzed the ability of the news and social media data to predict price fluctuations for three of the largest coins. They used text-based sentiment classification by first labeling the news and social media data instead of the traditional positive or negative method (3). Another from Dartmouth University used a similar sentiment-based approach, but focused on the gradient boosting tree model (4). Others have used time-series and other quantitative analyses (5), (6).

Our goal was to approach the problem differently to predict the cryptocurrency prices. We decided to focus our efforts on only four coins: Bitcoin, Bitcoin Cash, Ethereum, and Ripple. We utilized correlations and log transformations to modify variables for the best models. We mostly wanted to use linear regression and random forest models.

Data

We obtained a dataset from Kaggle that was originally taken from CoinMarketCap. The full dataset consists of daily data from April 2013 to July 2018 on roughly 1,350 cryptocurrencies for over 651,000 records.

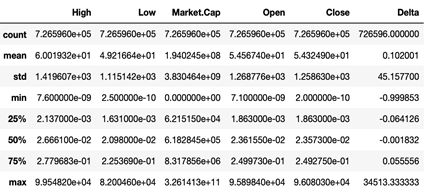
Data Description

Below is a list of the variables included in the dataset.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Description** | **Type** |
| Date | The day of recorded values | Date |
| Open | The opening price (in USD) | Numerical |
| High | The highest price (in USD) | Numerical |
| Low | The lowest price (in USD) | Numerical |
| Close | The closing price (in USD) | Numerical |
| Volume | Total exchanged volume (in USD) | Numerical |
| Market.Cap | The total market capitalization for the coin (in USD) | Numerical |
| Coin | The name of the coin | Categorical |
| Delta | Calculated as (Close - Open) / Open | Numerical |

Basic Statistics

Below is a quick summary of the data. Market Cap and Volume are in the millions whereas other price variables are in the thousands.



Because of the size of the dataset, we decided to only take the three most recent full years, from July 2015 to July 2018. This limited our dataset to just over 613,000 records to start.

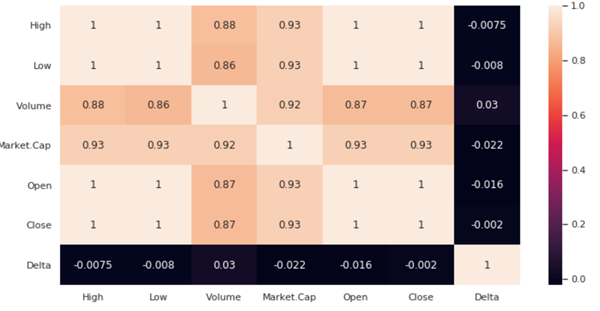
Empirical Results

From this initial dataset, we performed preprocessing to clean the data, transformations on certain variables, identified the coins to focus on, then ran a series of models before analyzing the results.

Data Exploration

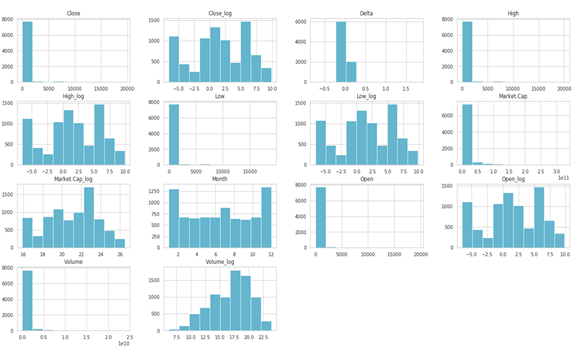
The first step we took was to clean the initial dataset to prepare it for analysis. Exploring the dataset for missing values showed that these records accounted for less than 1%. Given the small number, we decided to remove the records. In addition, we noticed that there was some invalid data in the Market Cap and Volume fields so identified and removed those records. These two columns were also string fields so had to convert to float type. To facilitate analysis, created new columns for Month and Year from the Date field. Found that some rows did not have a Close price (which was out Target variable) so removed those as well. For the remaining null values, filled in the Open and High variables with the average of each based on the Year. Using these, calculated the Delta. See Appendix 1 for the data preprocessing screenshots from Python.

With the cleaned dataset, we explored the data and the correlations between the variables. Below is the correlation matrix.

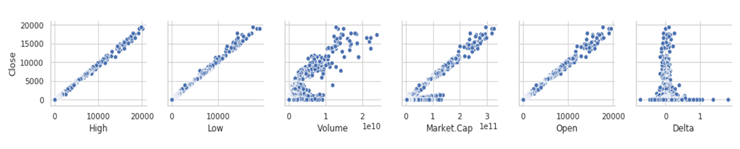


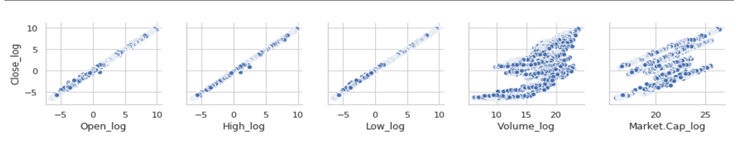
From the plot, we can see that Open, High, and Low variables all have a correlation of 1 with Close, our target variable. Market Cap and Volume have very high correlations with 0.87 and 0.93, respectively, with Close. The only variable that did not have a significant correlation with closing price was Delta (the relative change in open and close on a given day) so we decided to drop that variable from further analysis.

Furthermore, we looked at the distributions of each variable to consider the need for any transformation or standardization. Due to their high skewness, we decided to use log transformations on six of the variables: Close (target), High, Low, Market Cap, Open, and Volume. Below is a chart of the distributions for all variables and their log counterparts, if applicable. See Appendix 2 for boxplots of these variables after the log transformations.

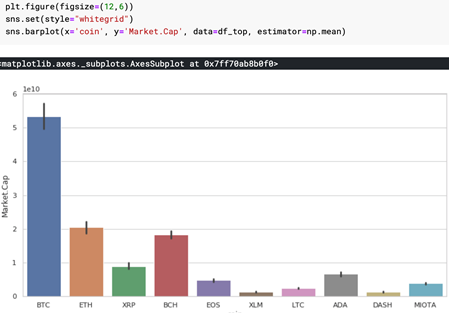


We found that after performing these transformations, the highly skewed distributions moved towards being less skewed. This would help us in understanding the effect of the change in the variables’ values on Close price. This is further evidenced in the scatterplot comparisons below. We see the correlations for Close vs. some other variables (no transformation) found in the top row of plots are very strong, but become even more apparent in the bottom row of plots with show Close vs. other variables (log transformed), particularly with High, Low, and Open. Notice that Volume and Market Cap still do not have a strong linear relationship even after the transformations.





Once we had the variables transformed, we wanted to filter the dataset to the top four coins. This would keep our analysis focused and most useful for findings and business impacts. Given our data from July 2015 to July 2018, we took the overall Market Cap value for each coin and found the top four were Bitcoin (BTC), Ethereum (ETH), Bitcoin Cash (BCH), and Ripple (XRP). Below is a plot of the distributions and a snippet of the Python code used.



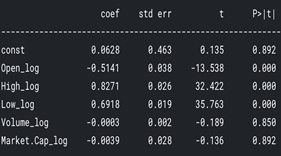
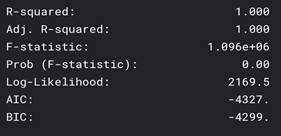
Bitcoin and Bitcoin Cash are two of the most popular and most expensive cryptocurrencies in the market with a relatively long presence. Ethereum and Ripple are newer and cheaper. Below is a chart of the closing price for each of the coins for the past year. Bitcoin had by far the highest market cap. In general, the market saw steep price increases in late 2018 followed by sharp price decreases until mid-year 2018. Since that time, the prices have been fairly stable and consistent.



Predictive Modeling

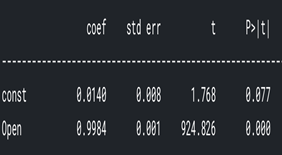
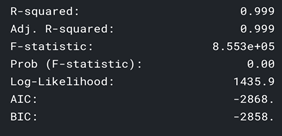
Once the dataset had been cleaned, transformed, and filtered to the top four coins of the market, we were ready to start the modeling process in predicting the closing price for each coin.

Since it is the predominant coin in the market, we decided to look at Bitcoin first. We performed a linear regression on the data with Close as the target and using the log transformed variables. The results are below.

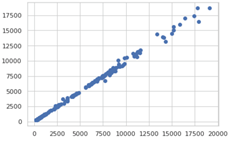
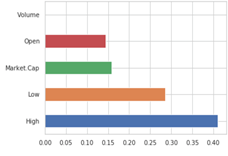


We see that while Volume and Market Cap are not significant for predicting given the low p-values for the coefficients, but the other variables are predictive. For example, 1% increase in High price would result in 0.83% increase in Close price. We can interpret the other variables similarly. We found this regression model with log transformed variables to be very accurate (r-square = 1.00) which could suggest we do not need a regression model to predict the closing price. Instead to just look at the High and Low prices of the coin. We also looked at the model if we standardized all the variables instead of using the log transformation. Those results were very similar and did not yield any noteworthy differences. See Appendix 3 for the results of that model.

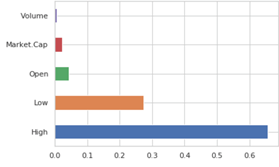
We decided to look at the regression model again, but using just the original Open variable (not transformed) as the predictor. The accuracy was still very high with r-square equal to 0.999. Results of the model are below. This gives us further evidence that a regression model might not be the best approach and decided to explore other model options.

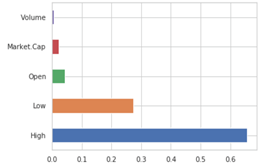


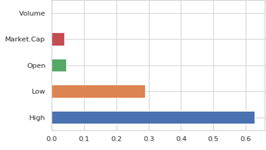
Continuing on to predict the closing price for Bitcoin, after regression we looked at a random forest model with all of the variables. For this approach, we used the original variables and not the log transformed ones. The left graph below is the feature importance plot which shows High, Low, and Market Cap as the most important variables contributing to the model. Similar to regression, the accuracy was very high (r-square = 0.998) and the model predictions were very close to the actuals. This is shown in the right plot below.



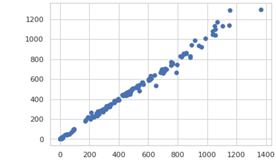
Given the strength of the random forest model for Bitcoin, we decided to use that model for the remaining three coins: Ethereum, Bitcoin Cash, and Ripple. We found that for each coin, the variables that were the biggest predictors for Close were High, Low, and - to a lesser extent - Open. Below are the feature importance plots for each coin in the order of Ethereum, Bitcoin Cash, and Ripple from first to last.

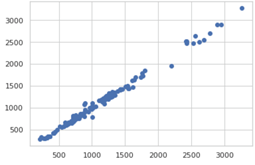


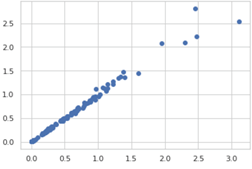




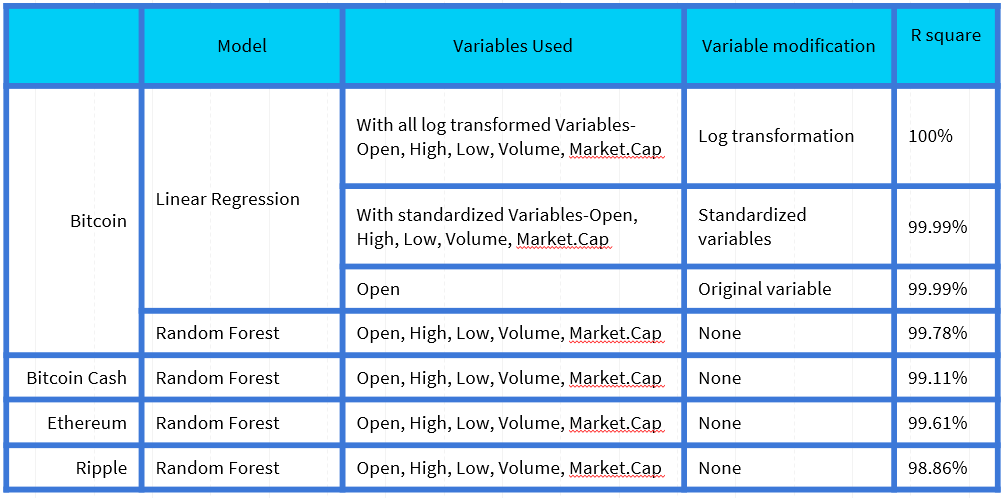
Similar to the Bitcoin model, each of these models yielded high accuracy results. The r-square values were 0.996, 0.991, and 0.989 for Ethereum, Bitcoin Cash, and Ripple, respectively. The scatterplots below show the predicted vs. actuals for each model in the order of Ethereum, Bitcoin Cash, and Ripple from first to last. We see that there is a strong linear relationship between the modeled and actual results, giving further evidence of the model’s accuracy and strength.







Below is a summary of the different models run for each of the coins, the variables used, any modifications to these variables, and the r-square values.



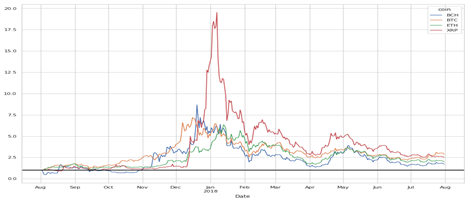
Findings

We found that both linear regression and random forest regressor models perform well in predicting the closing price for the top four coins. The closing price depends significantly on other pricing variables of High and Low as well as Open to a lesser degree. Even using a single variable, Open, on a Linear Regression model for Bitcoin, we yielded high accuracy results.

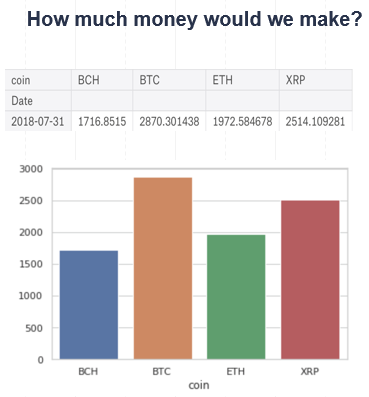
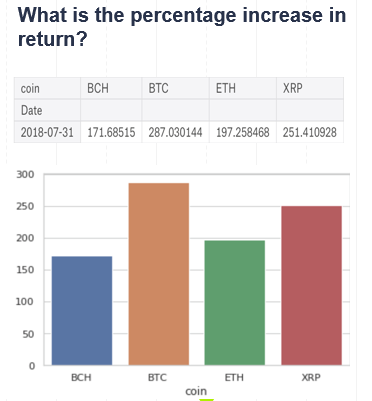
Implications/Recommendations

Given these results, a model might not be needed for predicting the closing price of these coins. Rather than implementing a model, investors and companies alike can watch the High and Low values to effectively estimate the closing price of a coin for that day. In addition, the Open price also gives a fair estimate of what to expect by day’s end.

With these coins, it might be difficult to determine which one to invest in. Below is a chart of the returns for each coin over the last year. Looking at the individual coins, Ripple has the highest return compared to the others.



Furthermore, we can look at the percentage increase on return and the net money gained from investing in each coin. Those results are below and we see that Bitcoin and Ripple yield the best results.



With the coins performing relatively close, we looked to see what we could invest in using only $1,000. Below is a table and a chart showing the number of units for each coin we can buy. We see that we can purchase less than 5 units for Bitcoin, Bitcoin Cash, and Ethereum, but over 5,700 units for Ripple.



We would recommend companies and investors to look at Ripple in terms of cryptocurrency investment. Currently, it is the cheapest coin out of the four and has an increasing market trend so there seems to be a promising outlook to see a high return on Ripple. While a model might not be needed, it is important to look at the High and Low price for the coin on a given day as well as the Open price to determine the Closing price and yield the best returns.

References

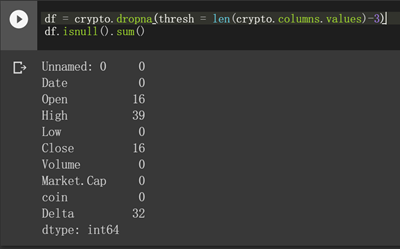
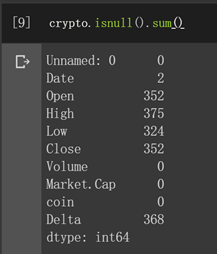
1. Martucci, B. (2019). What Is Cryptocurrency - How It Works, History & Bitcoin Alternatives. Retrieved from https://www.moneycrashers.com/cryptocurrency-history-bitcoin-alternatives/
2. Reiff, N. (2019). Bitcoin-Related Stocks Aim to Capitalize on Cryptocurrency Craze. Retrieved from https://www.investopedia.com/news/bitcoinrelated-stocks-aim-capitalize-cryptocurrency-craze/
3. Lamon, C., Nielsen E., & Redondo E. (2017) *Cryptocurrency Price Prediction Using News and Social Media Sentiment*. [PDF]. Retrieved from https://pdfs.semanticscholar.org/c3b8/0de058596cee95beb20a2d087dbcf8be01ea.pdf
4. Li, T., et al. (2019). *Sentiment-Based Prediction of Alternative Cryptocurrency Price Fluctuations Using Gradient Boosting Tree Model.* Retrieved from https://www.frontiersin.org/articles/10.3389/fphy.2019.00098/full
5. Catania, L., Grassi, S. & Ravazzolo F. (2018) *Predicting the Volatility of Cryptocurrency Time-Series*. Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-89824-7\_37
6. Saad, M. & Mohalsen, A. (2018) *Towards characterizing blockchain-based cryptocurrencies for highly-accurate predictions.* Retrieved from https://ieeexplore.ieee.org/abstract/document/8406859

CoinMarketCap. https://coinmarketcap.com/charts/

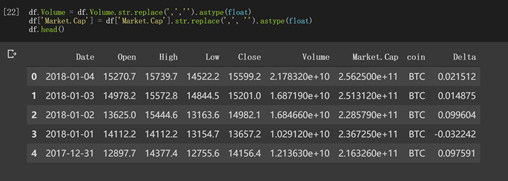
Kaggle. https://www.kaggle.com/

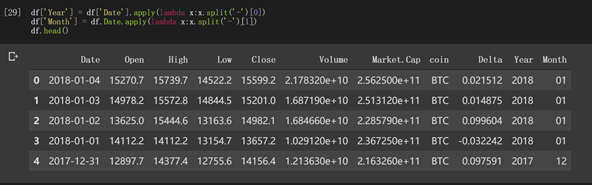
Appendix

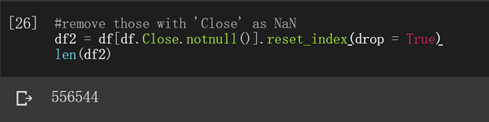
Appendix 1 - Data Preprocessing

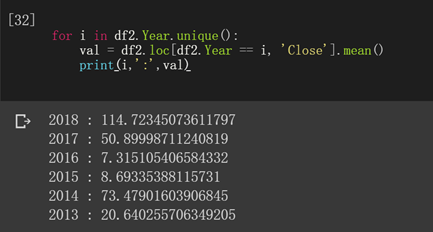


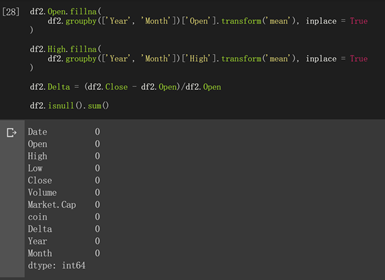




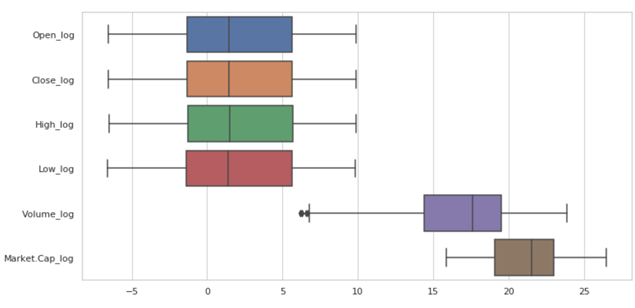








Appendix 2 - Boxplots of Log Transformed Variables



Appendix 3 - Results of Linear Regression Model for Bitcoin Using Standardized Variables

