DATA MINING AND PREDICTIVE MODELING FOR CRYPTO-CURRENCIES

[Progress Report]

James Beedle
University of Colorado
Boulder, CO USA
James.Beedle@Colorado.edu

Lauren Deans
University of Colorado
Boulder, CO USA
Lauren.Deans@colorado.edu

Inderpal Dhillon
University of Colorado
Boulder, CO USA
indh2640@colorado.edu

Theo Shin
University of Colorado
Boulder, CO USA
Ted.Shin@colorado.edu

KEYWORDS

Data Mining, Cryptocurrency, Machine Learning, Linear Regression, Predictive Modeling, Python, Bitcoin, Ethereum, Litecoin, Ripple

ACM Reference format:

James Beedle, Lauren Deans, Inderpal Dhillon and Theo Shin. 2019. Insert Data Mining and Predictive Modeling for Crypto-Currencies: Proposal Paper. *University of Colorado, Boulder, CO, USA, 4 pages.*

1. PROBLEM STATEMENT

With over one million data points structured to varying bit-currencies, our objective is to discover correlations involved with forecasted predictions and prevailing trends. We have begun our deep dive into the data and have already discovered interesting insights. Through data cleansing, our group will be focusing on aggregating crypto data through prior years and following the removal of incomplete data, handle the outliers and noisy data. Our preliminary pre-processing efforts will lead to a complete integration containing the additional datasets from alternate crypto-currencies, in order to simplify our analyses and form prediction models.

Questions and deeper meanings our group seeks to answer include which currency provides the best average returns for investments, which trends are consistent amongst various groups/countries, any unique correlations which can potentially be linked to our predictions, and which method can yield the most accurate forecast. Similarly, another question we would like to answer is does a Bitcoin dollar increase necessarily lead to increases across the basket of currency indices? Our hypothesis is that the alternate cryptocurrencies may even have a negative relationship compared to Bitcoin.

Once correlations are discovered, we would like to understand the reason why these correlations exist and determine how these trends came about. Which formulated prediction models can be the most accurate in the future and seeing how correlations are tied to our data analysis.

2. LITERATURE SURVEY (PREVIOUS WORK)

Bitcoin came onto the scene as the first decentralized cryptocurrency and was first released as open-source software in 2009 (created by alias Satoshi Nakamoto [1]) and since, there have been over 4,000 alternative variants of cryptocurrency in the market. Bitcoin has grown tremendously since then with prices less than 15 USD in 2012 to its current market value at 8122 USD (as of Oct 24, 2019). The historical high for Bitcoin occurred on Dec 14, 2017 at a value near $18,000 USD [2].

Coincidentally, data mining tools and methods have improved within the same timeframe which has encouraged many individuals to attempt prediction models and figure out when the next jump in price could take place.

Cryptocurrencies have become a popular form of transaction amongst merchants and investors for multiple reasons. To understand why more and more

people are using cryptocurrencies for market exchange, let's look at the advantages:

- Not needing a middle man (no financial institutions and no transaction fees)
- Anonymous transactions and omission of personal details
- Facilitating International transactions (Bitcoin is not tied to any country's regulations)
- Investment opportunity (If one had invested $100 in Bitcoin in its early stages the investment would be worth north of $25 million).

The uncertainty and volatile nature underlying Bitcoin has created curiosity and even concern, as no one really knows what will become of Bitcoin. Looking at the pattern for the past 2 years, the constant fluctuation in the price of Bitcoin is a challenge for prediction. Despite its unevenness, Bitcoin's value still remains relatively high.

Figure 1: Five-year trend of Bitcoin value [2].

What our group would like to do is replicate and even continue the predictive analysis for cryptocurrencies using various data mining techniques. An applicable machine-learning model we are initially implementing are Time Series models to examine past patterns and trends to anticipate future fluctuations.

3. PROPOSED WORK

In order to begin the data collection and analysis, we must first pre-process the dataset. Using historical bitcoin data acquired from https://www.kaggle.com/albala/ticks-bitcoin-ethereumlitecoin-ripple, we begin by formatting the CSV data into a Pandas DataFrame to facilitate exploratory data analysis (EDA) adding plots and calculations. We aim to back the accuracy of our findings using statistics. Plots will be developed in order to visualize price movements in our DataFrame between the various cryptocurrencies over a set amount of time. This will involve adding a numerical tag to categorize and describe the cryptocurrency type:

1. Bitcoin cash
2. Bitcoin
3. Ethereum
4. Litecoin
5. Ripple

Since each alternate cryptocurrency is stored in separate CSV files, we will combine all into a single list totaling 1,314,830 rows of data points. This will invariably lead to cleaning the data (null/missing, redundancies, and removal of noisy data). To account for redundancies, our plan is to remove repeatedly occurring data that does not aide in our analysis and predictions. For missing data, we will have to determine values to drop or replace with the most applicable statistical measure. By choosing to replace missing data with the mean or other method, we can minimize any negative impact towards our conclusions and insights.

After dropping the categories 'order number,' 'buy order number,' 'sell order number,' and 'type' columns, the data needs to be rearranged and grouped by correlations (chronological, closing costs) and possibly even develop a new attribute type such as average closing costs, to increase efficiency in our analysis and prediction models and to make clearer sense of the data. Throughout our analysis, we will keep track of which correlations are prevailing and which form of cryptocurrency is directly influenced by others. If Bitcoin increases by a significant measure, does that necessarily mean Ethereum or Litecoin necessarily increase? And by how much?

As previously mentioned, prediction forecasting for cryptocurrencies has been developed and studied in the past. Many individuals seek to predict not only cryptocurrencies, but the stock market, housing prices, and other investment vehicles to discover financial opportunities. Some of the past work have aspects of technical work beyond the scope of our project (TensorFlow, Keras, Deep Learning), however, it is interesting to glean into one of many possibilities using data science practices [6].

A statistical implementation we feel is important to our analysis is the Dickey-Fuller statistical test, which implements a stationarity and thus removing possible intrusiveness to our dataset and allowing for better performing prediction models. With the Dickey-Fuller test, we will be aiming at P-Values less than the critical 5% and attempt at lowering P-Values as close to zero as possible for a high measure of accuracy [3].

Our project will entail observing the historical data associated with Bitcoin and four other alternates, which should provide interesting correlations or distinctions between the datasets.

We will also be using multilinear regression models to determine any correlations that exist. We will be utilizing a pair plot to visualize the multilinear regression, as well as a correlation table.

If time permits, we will also be exploring other models such as Support Vector Machine (SVM) models, which are supervised learning classifiers that will aid in the future predictions of the cryptocurrencies using the provided dataset to train the model.

4. DATASET

The dataset includes a collection of five different cryptocurrencies along with their transactions. The datasets start on 8/15/18 and end on 9/22/18 inclusively, for a total of 39 days of acquired data with a total of 1,314,830 data points. The dataset includes the following four attributes:

- Trade price in USD
- Transaction date and time
- Coin valued amount per transaction
- Type of Coin

We found our dataset from https://www.kaggle.com/albala/ticks-bitcoin-ethereumlitecoin-ripple [4], though the site states it was originally aggregated from 'Coinbase.' The dataset is downloaded and accessible using our personal computers, as well as stored online in our GitHub repository. In CSV format, the dataset is 50.7MB in size. Since our datasets are static in nature, our analysis should be accurate so long as we mitigate any intrusive trends (using Dickey-Fuller) and preserve our relative data points following cleaning and pre-processing.

Our dataset has been converted to a matrix which will allow for easier comparisons and analysis of the different cryptocurrencies based on a similar time series of transactions. The data is arranged based on the time series attribute, which allows for direct comparisons based on date and time of a transaction. The time stamp has been truncated to eliminate seconds and simply include time stamp up to the minute.

5. EVALUATION METHODS

Predicting BitCoin and alternate cryptocurrency prices using time series forecasting can be considerably different from the usual machine learning models, mainly because of its time dependency. The basic assumption of using a linear regression model that observations are independent won't apply for our application. Since simple machine learning models won't work for some of our prediction analysis, past work has used article time series models such as Auto Regressive (AR) model, Moving Average (MA) model, and Autoregressive Integrated Moving Average (ARIMA) model which can all be used for forecasting Bitcoin price.

As previously discussed in our proposed work, we will look to test for stationarity using the Dicky-Fuller statistic test to determine how strongly a time series is defined by a trend. To test, the Null Hypothesis (H_0) can be represented as the time series being not stationary and alternatively, our H_1 hypothesis will state the time series is stationary. As already

mentioned, we can reject the Null Hypothesis if P-Values are determined to be less than or equal to 0.05 and conclude the data is stationary [8].

We expect to be challenged with highly skewed data (due to the volatile trends of Bitcoin) and will most likely need to transform the series (log transforming and/or differencing), in order to remove intrusive trends and increase chances of stationarity and lowering the P-value to 0.05 or below [8]. These transformations should help in our forecasting process by stabilizing mean values over time.

Of the previously mentioned time series models, the ARIMA model looks to be the most promising for our application as demonstrated in previous work to minimize the residual sum of squares error and provide reliable predictions. For our prediction model, we will split train and test data and test for mean error differences between predicted and expected values calculated (error = |predicted – original)/ (original * 100).

We will also be conducting data analysis using a multilinear regression model. This will allow for exploration of correlation between the different cryptocurrencies but stops short of allowing any predictive outcomes. The multilinear regression will be utilized to search for correlation between the data, with other models implemented for predictive measures extending into the future. To visualize these potential correlations, a pair plot will be generated.

We will also be exploring supervised learning models such as support vector machines (SVM) which will supplement the multilinear regression in that it is capable of producing predictive measures that extend into the future which will allow us to determine and predict the future prices of these cryptocurrencies given our dataset used as a training set for the model.

6. TOOLS

Python offers extensive access to built-in libraries and modules which will be utilized in our project that are especially useful for financial analysis and linear regression models. Specifically, we be utilizing modules such as seaborn through matplotlib, sklearn, pandas, numpy, nltk, scipy, scipy.stats, candlestick, and wordcloud. We will be creating a pandas dataframe containing cleaned and organized data which will then be converted to a numpy array, and then further our prediction model which will utilize support vector machines with 80% training and 20% testing. Most of this analysis will be conducted within a Jupyter notebook. Microsoft Excel's currency converter will be utilized when necessary to relate each cryptocurrency to one another and relative to a USD. We will be conducting analysis using statistical methods such as Naïve Bayes classification models, as well as the machine learning task of supervised learning. We may look to host our data using a relational database system such as MySQL to gain insight over the data and allow additional exposure to SQL querying.

7. MILESTONES

In accordance to deadlines set by the class, we will also implement the following milestones to ensure adequate progress towards completion:

- November 1- Data cleaned, pre-processed, derived attributes integrated into dataset.
- November 5- Generate hypothesis for which statistical algorithm will most accurately forecast trends.
- November 8- Complete code for statistical analysis of complete dataset within Jupyter Notebook, generate code for tracking trends/ correlations between the different cryptocurrencies.
- November 12- Apply statistical analysis methods to data, finalize algorithms to incorporate for predictive forecasting models.
- November 15- complete part 3 (Progress Report).
- November 23- Apply forecasting models and determine which method produces most accurate results relative to actual trends.

- December 4- Complete statistical analysis of correlations between different cryptocurrencies.
- December 9- Finalize results and complete visualization, complete presentation.
- December 13- Ensure Completion of parts 4-7 (Final report, code and descriptions, presentation, peer evaluation and interview).

7.1 MILESTONES COMPLETED

At the time of writing, a new problem has been determined which will require restructuring of our milestones to ensure adequate completion of the project. Once the data was transformed into usable CSV format, it consisted of 1,314,830 data points, each in individual rows based on transaction time and type of currency. Each individual transaction was displayed in its own row. When beginning to manipulate and explore our data, we realized that the way the data was displayed proved to be difficult when conducting analysis. This resulted in a decision to change our data format, which has now been converted to a data matrix. Using a unique list of dates and times, each coin was placed in the matrix based on its sale and price, since there were instances of many trades occurring within a minute, the last price within a minute is the total price. This effectively condensed down our data and will allow for easier analysis as the condensed matrix is much more manageable.

Once the data was converted to the matrix format, a problem arose in that the multilinear regression model did not prove to provide accurate results based on the previously unknown null values presented in the matrix as 0.0. This skewed the analysis and has proven as something that needs to be addressed before any further analysis can be conducted. A pair plot was generated (see Figure 2) which highlighted the need to remove or account for the zero values to accurately assess correlation between the different cryptocurrencies.

7.2 MILESTONES REMAINING

The milestones remaining will be in accordance to deadlines set for the class to ensure timely and thorough completion of the project. The milestones have been adjusted and restructured to allow for the necessary changes to be made and data manipulation to occur to determine which format will allow for the data to be best and most easily analyzed. The remaining milestones are outlined below:

- November 18- Adequately account for missing transaction values, either using 'NaN' values or utilizing mean of nearest transactions for a given cryptocurrency to fill in the missing/inadequate data values.
- November 21- Apply multilinear regression models, Dickey-Fuller statistical analysis, and utilize trained data using support vector machines (SVM) model with supervised learning methods.
- November 23- Determine which analysis method produces most accurate results based on actual data trends in our cryptocurrency dataset.
- December 4- Complete statistical and data analysis and reach predictive conclusions for the cryptocurrencies.
- December 9- Finalize results, produce and complete an HTML page to display analysis results visually.
- December 11- Complete and finalize presentation and final report.
- December 13- Ensure adequate completion of parts 4-7 of the project, which include the final project report, code and descriptions, presentation, peer evaluations, and interview.

8. RESULTS PROGRESS

The results generated so far have surfaced the need to restructure the data to account for the missing transaction values. This problem was not known to us until further work was conducted using the dataset and the dataset was restructured into a matrix format to

allow for easier and more manageable use of the dataset. A pair plot was generated using the dataset in the new matrix format, which can be seen below. The results of conducting this pair plot allowed us to easily visualize the issue of the zero values within the data matrix. The generated pair plot can be found below (Figure 2).

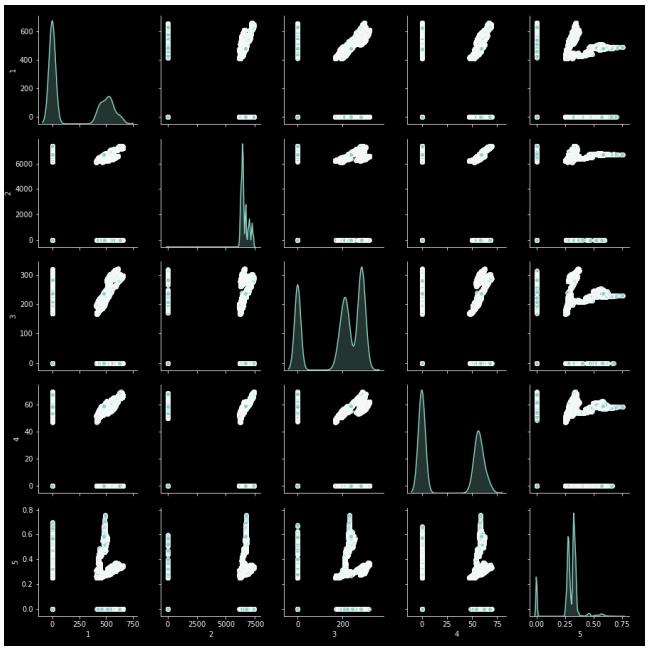

Figure 2: Initial generated pair plot comparing the five different cryptocurrencies utilized in our project.

As you can see, the results do not provide any sort of insight due to the zero data values, able to be visualized along different x and y axes within the pair plot. Once this data has been cleaned to account for the zero values, the pair plot will be generated once again and will produce results that provide much richer insight to correlations between the different cryptocurrencies.

The multilinear regression model was also conducted using the initial regenerated dataset matrix and provides similar insight as the pair plot above. The initial linear regression table can be found in Figure 3.

OLS Regression Results

Dep. Variable:	Q("1")	R-squared:	0.077
Model:	OLS	Adj. R-squared:	0.077
Method:	Least Squares	F-statistic:	1124.
Date:	Wed, 13 Nov 2019	Prob (F-statistic):	0.00
Time:	18:43:31	Log-Likelihood:	-3.7358e+05
No. Observations:	54053	AIC:	7.472e+05
Df Residuals:	54048	BIC:	7.472e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	19.8937	4.955	4.015	0.000	10.181	29.606
Q("2")	0.0098	0.001	16.164	0.000	0.009	0.011
Q("3")	0.3056	0.009	32.945	0.000	0.287	0.324
Q("4")	1.7239	0.037	46.676	0.000	1.651	1.796
Q("5")	44.9234	10.000	4.492	0.000	25.324	64.523

Omnibus:	280721.386	Durbin-Watson:	1.277
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6103.279
Skew:	0.499	Prob(JB):	0.00
Kurtosis:	1.691	Cond. No.	6.38e+04

Figure 3: Initial linear regression table.

As you can see, the R^2 value is very small in the linear correlation table, which in practice can be interpreted as the data should not be considered for correlation at this point. Further manipulation of the data (removal of zero values) will prove to change this R^2 value to a hopefully higher value.

REFERENCES

[1] Wikipedia. 2019. Wikipedia: The Free Encyclopedia. Retrieved from http://en.wikipedia.org/wiki/Bitcoin.

[2] Refinitiv. 2019. Retrieved from http://Refinitiv.com

[3] Bhavesh Bhatt. 2019. Adf-test-stationarity-python. (October 2019). Retrieved October 24, 2019 from https://github.com/bhattbhavesh91/adf-test-stationarity-python/commits/master/augmented-dickey-fuller-test-python.ipynb.

[4] DataJuicers. 2018. DataJuicers cryptodatabase: bitcoin, ethereum,litecoin, ripple. (September 24, 2018). Retrieved on October 24, 2019 from http://www.kaggle.com/albala/ticks-bitcoin-ethereumlitecoin-ripple.

[5] Wikipedia. 2019. Wikipedia: The Free Encyclopedia. Retrieved from http://en.wikipedia.org/wiki/Cryptocurrency.

[6] Cryptocurrency Price Prediction Using LSTMs | TensorFlow. 2019. Retrieved from https://towardsdatascience.com/cryptocurrency-price-prediction-using-lstms-tensorflow-for-hackers-part-iii-264fcdbccd3f.

[7] Machine Learning Mastery. 2019. Retrieved from https://machinelearningmaster.com/arime-for-time-series-forecasting-with-python/.

[8] Stationarity and Differencing. 2019. Retrieved from https://people.duke.edu/~ma/411diff.htm.