Analysis of Bitcoin, Ethereum and Litecoin using Apache spark Machine Learning Library

Course 3252: Big Data Management Systems and Tools
University of Toronto
Kushagra Dixit

Introduction

Crypto-Currency is a digital asset designed to work as an exchange medium, utilizing strong cryptography to secure financial transactions and to control the creation of additional units for currency value moderation and to verify asset transfer. Cryptocurrencies employ decentralized control as opposed to centralized digital currency and central banking systems. The decentralized control of each cryptocurrency typically works through a blockchain that also serves as public financial transaction database.

According to Jan Lansky, a cryptocurrency is a system that meets the following conditions:

1. The system does not require a central authority and its state is maintained through distribution consensus.
2. The system keeps an overview of cryptocurrency units and their ownership
3. The system defines whether new crypto-currency units can be created and the circumstances of the origins of new units and how to determine the ownership of new units.
4. Ownership of cryptocurrency can be proved exclusively with cryptography.
5. The system allows transaction to be performed in which the ownership of the cryptocurrency units can be changed and a transaction statement can only be issued by an entity proving the current ownership of the assets
6. If two different instructions for changing the ownership of the same cryptographic units are simultaneously entered, the system performs at most one of them.

Fulfilling these conditions, Bitcoin was invented by an unknown person or group of people using the name “Satoshi Nakamoto”. The Bitcoin network was created on January 3rd 2009 when Nakamoto mined the first block of the chain, also known as the genesis block. Bitcoin was then first released as an open source software in 2009 is still considered the first decentralized cryptocurrency. The ticker symbols used to represent Bitcoin are BTC and XBT and its Unicode character is ₿. Smaller alternative units of bitcoins include milli-bitcoin (mBTC), and “satoshi” (sat), where satoshi named after the bitcoin’s creator is equal to 0.00000001 bitcoins.

Bitcoins were created as a reward for a process known as mining. Mining is a record keeping service done through the use of computer and graphic card processing power. Miners keep the blockchain consistent, unalterable and complete by repeatedly grouping newly broadcast transactions into a block which is then broadcast to the network and verified by recipient nodes. The bitcoin blockchain is a public ledger that records any transactions performed by using bitcoins. The Bitcoin blockchain can be explained as being a chain of blocks containing a hash of the previous block up till the genesis block of the chain. A simplified diagram of a bitcoin blockchain can be seen in figure 1.1 on the right.

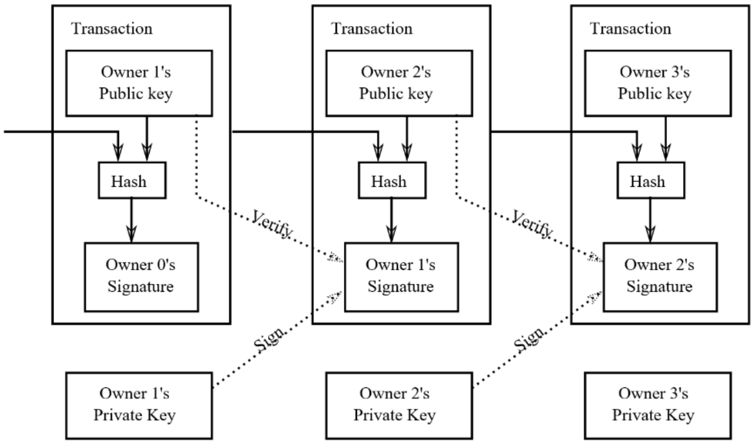
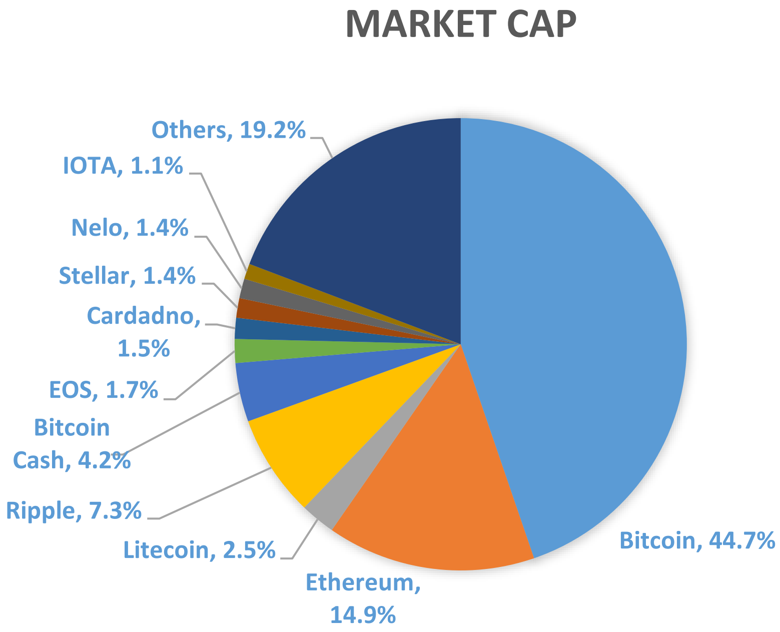


Figure 1.1: Simplified Bitcoin Blockchain

Research by University of Cambridge estimates that in 2017, there were 2.9 to 5.8 million unique users using a cryptocurrency wallet with most of them utilizing Bitcoin. Since the release of Bitcoin, over 4000 alternative cryptocurrencies have been created. Cryptocurrencies such as Ethereum, Litecoin, Ripple, EOS etc occupy the leftover 55.3% of Market CAP as an alternative to Bitcoin.



Ethereum on the other hand was created as a public, open source blockchain based distributed computing platform featuring smart contract functionality. Its ticker symbol is ETH and the Greek uppercase Xi character (Ξ) is generally used for its currency symbol. It was proposed in late 2013 by Vitalik Buterun, a cryptocurrency researcher and programmer by using online crowd-sale that took place between July 2014 and August 2014. 72 million coins were already mined when the system went live on 30th July 2015. The blockchain generated by the Ethereum platform is an "Ether". The smallest unit of Ether is "wei" which is defined as 1/10^{18} of an Ether. The Ethereum Virtual Machine or EVM utilizes international network of public nodes for executing scripts. This virtual machine and its instruction set is in contrast to the Bitcoin Script.

Ethereum also has some other significant differences when compared to Bitcoin, its block time is 14-15 seconds which is much faster when compared to 10 minutes for bitcoin. Its transaction fees differ by computational complexity, storage and bandwidth needs whereas bitcoin transactional fees is determined by the transaction size in bytes. The transactional fees for Ethereum is determined by the above stated variables using a system called "gas". Each of the "gas" units have a price that is specified in a transaction. This is measured in "wei" whereas Bitcoin transactions have fees specified as satoshis per byte. Furthermore, Ethereum debits and credits values in Wei from an Ethereum account system, as opposed to Bitcoins UTXO system which resembles spending cash and receiving change in return.

Similar to Ethereum, Litecoin was also developed as an alternative to Bitcoin. It is very similar to bitcoin when it comes to its technical details and was developed relatively early in October 2011 via an open source client on GitHub by Charlie Lee, a google employee and former engineering Director at Coinbase. Its ticker symbol is LTC and the letter Ł. It was divergent from the Bitcoin Core client by having a decreased block generation time of 2.5 minutes, a higher number of maximum coins, a modified GUI and different hashing algorithm scrypt, as compared to bitcoin's SHA-256. The utilization of scrypt which is a memory-hard function requiring more memory in its proof-of-work algorithm, the devices used for mining Litecoin are more complicated to create and more expensive to produce than they are for Bitcoin.

Objective

Various average pricings and transactional Volume for Bitcoin, Ethereum and Litecoin in a given day are the major topics of analysis in the report. The free and open to research dataset obtained through <www.cryptodatadownload.com> provides insight into the Opening and Closing prices of the three cryptocurrencies. Since, cryptocurrencies generally operate on a 24 hour index, the date and time assumed for the closing and opening of these values is 0:00hrs for the day. The dataset's also list the High and Low Price points of the cryptocurrency in a given day, and the cryptocurrencies Transactional volume in US dollars (USD) and in other cryptocurrencies (BTC, ETH and LTC)

To analyze the datasets given the information above, multiple criteria was used to see the strongest determinants of Average pricing and Transactional volume of the three currencies. To use average pricing as an independent variable, it was calculated from the High and Low pricing for the day. To index the content of the report, the following is a list of analysis that were performed using Apache Spark's Machine Learning Library (MLLib) to perform regression analytics by training on 75% of the dataset and applying the regressed model on the remaining 25% of the dataset:

1. How well do the high and low prices determine the transactional dollar volume in USD for BTC, ETH and LTC
2. How well do the opening, closing and transactional volume in USD of the day determine the average cost for the 3 crypto-currencies.
3. How well does Bitcoin open, close and average and transactional volume in USD determine the average cost of Ethereum and Litecoin
4. How Ethereum open, close and average and transactional volume in USD influences Litecoin average price
5. How Litecoin open, close and average and transactional volume in USD influences Ethereum and average price.

Some of the common metrics used to assess the regression analysis will be the mean squared error (MSE) of the regressed model to measure the average of the squares of the errors. Root mean square value will be used to explain the deviations between values and predictions by the regressed model. These deviations are also called residuals when the calculations are performed over the sampled dataset. To expand on the value and to determine the models ability to predict smaller variation vs large variation of errors in its dataset, a fitted vs residuals graph will also be employed. Mean Absolute Error (MAE) will be evaluated to measure the difference between two continuous variables for each model. To further test the fitting the regressed model and the proportion of variance in the predicted variable that is from the independent variable in the dataset, the coefficient of determination or R^2 value will also be used. Moreover, the "explained variance" that measures the proportion to which the regressed model accounts for the variation (dispersion) of a data set under consideration will also be utilized to demonstrate the quality of the regressed model. A prediction vs actual values will also be used to demonstrate the effectiveness of the regressed model.

Other than pure statistical analysis of the regressed model and its predicted values, The dataset will also be qualitatively evaluated for certain specific sections of the BTC, ETH and LTC average price and transactional volume in USD such as the slow but consistent rise of cryptocurrency value from 2011-2016, the December 2017 cryptocurrency bubble and subsequent crash and the subsequent recovery and its aftermath. Since, the evaluated pricing of cryptocurrency replies on its cost as decided by the market value, there is always uncertainty to be accounted for due to the volatile nature of the market.

Results

The effect of variation in High and Low daily prices and its effect on Bitcoin, Ethereum and Litecoin

Bitcoin:
The value of High and Low prices per day as an indicator for transactional Volume in USD for Bitcoin was used in a machine learning regression producing the following Prediction vs Actual Values graph as seen below in Figure 2.1

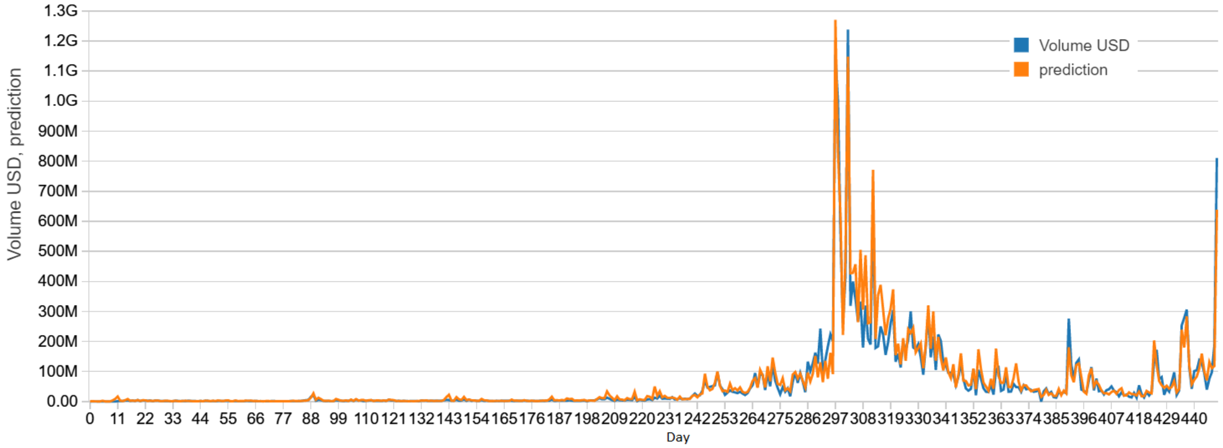


Figure 2.1: Predictor of Volume USD using High and Low pricing of Bitcoin

The fitted model above generated comparatively high RMSE of 3.155e7, and a MAE of 1.448e7 and an explained variance of 1.688e16 which demonstrate significant deviation and difference in the predictive datasets error calculation when compared to the actual Volume USD. This primarily could be attributed to almost 90% of the dataset being very small compared to the two massive peaks that occur around day 300 of the tested model. The high and low pricing around day 300 of the test model marks around Bitcoin bubble period where the transactional Volume in USD was jumped from 40,000-50,000,000 times the price of Bitcoin. An R^2 value of 0.937 communicates that the predicted model is statistically very significant. However the model is not without its limitations, as the ratio between the variables of transactional volume and daily pricing increases, the fitted vs. residuals plot shows heteroscedasticity in our regressed model. This can be attributed to the unstable nature of Bitcoin especially when the cost and transactional volume balloons way higher compared to 90% of the selected dataset.

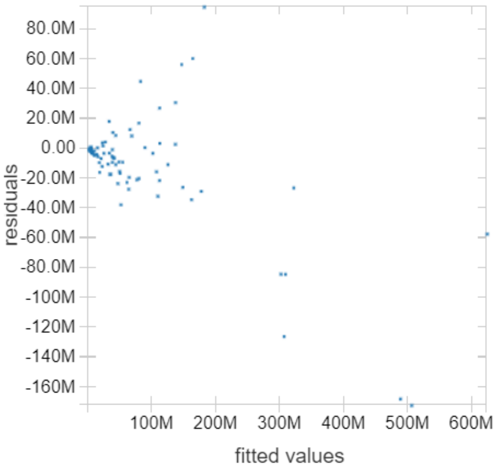
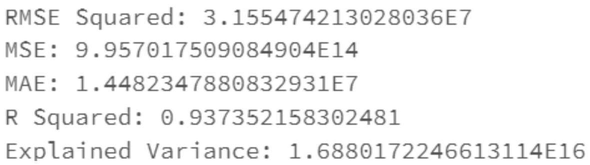


Figure 2.2: Residual vs Fitted for Bitcoin transactional Volume in USD



```
RMSE Squared: 3.155474213028036E7
MSE: 9.957017509084904E14
MAE: 1.4482347880832931E7
R Squared: 0.937352158302481
Explained Variance: 1.6880172246613114E16
```

Ethereum:
Similar to how the data was regressed and modelled for Bitcoin, the data of high and low prices to predict the transactional volume of Ethereum demonstrated the following prediction plot:

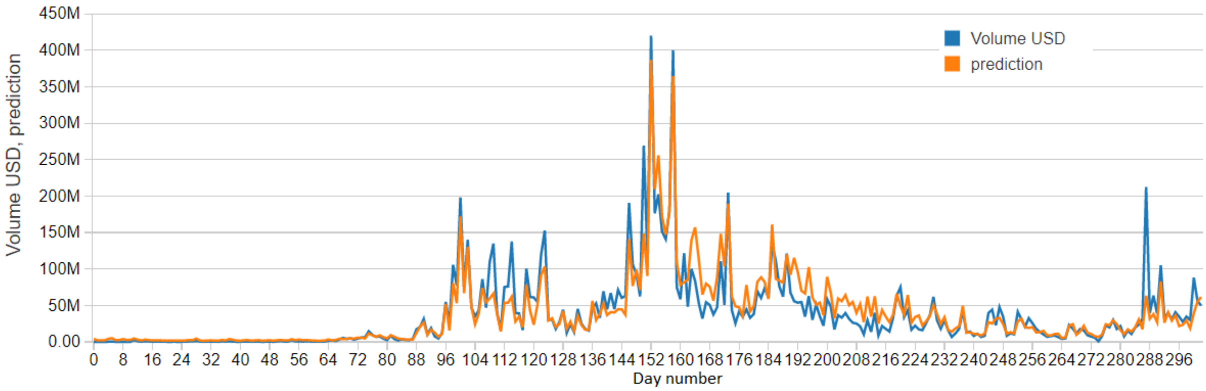
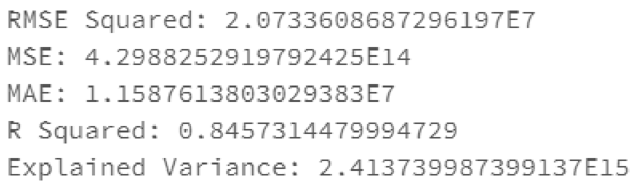


Figure 2.3: Predictor of Volume USD using High and Low pricing of Ethereum

The regressed model for ETH generated similar but comparatively lower Root mean square deviation of 2.0733e7 and Mean absolute error of 1.158e7 showing lesser deviation in the predictive datasets error calculation when compared to the values of Bitcoins predictive Dataset. Despite, the decrease in error calculation, a lower R^2 value of 0.845 and explained variance of 2.41e15 shows that that predicted model is statistically significant, (as R^2 is >0.7) but is a weaker fit when compared to the Bitcoin's predictive model in Figure 2.1. This is primarily due to ETH dataset having more minor peaks causing difficulties in predicting the Transactional Volume variation. Similar to Bitcoin's dataset, as the values go higher, as in the case of the two major peaks occurring that the December 2017, the associated risk increases tremendously. The margin of error as seen in the fitted vs. residuals plot shows lesser heteroscedasticity when compared to Bitcoin. This can be attributed to the similar unstable nature of Ethereum especially when the transactional volume becomes 5-10 times of almost 80% of the selected dataset


```
RMSE Squared: 2.0733608687296197E7
MSE: 4.2988252919792425E14
MAE: 1.1587613803029383E7
R Squared: 0.8457314479994729
Explained Variance: 2.413739987399137E15
```

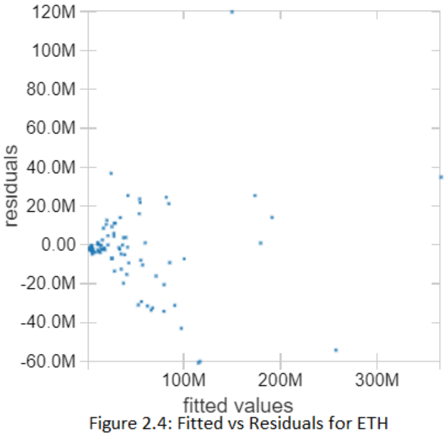


Figure 2.4: Fitted vs Residuals for ETH transactional Volume USD

Litecoin:

When compared to Bitcoin and Ethereum, Litecoin generates the following modelled fit for its Transactional Volume in USD

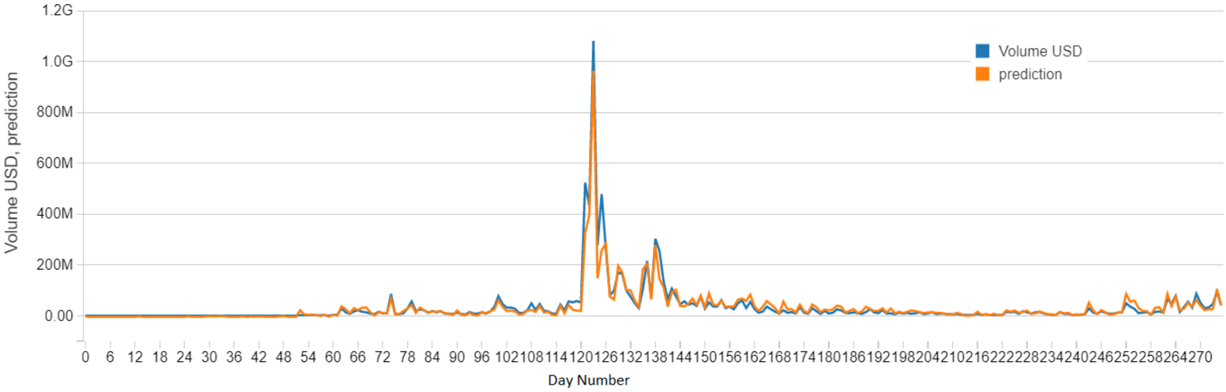
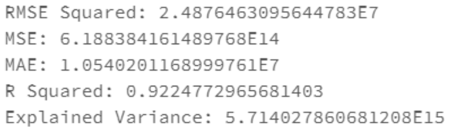


Figure 2.5: Predictor of Volume USD using High and Low pricing of Litecoin

Litecoin’s transactional volume regression above had lower comparative Root mean square deviation of 2.487e7, which is higher than Ethereum but lower than Bitcoin. This was primarily because of high stability and lesser number of smaller sized peaks. However, the Mean absolute error of 1.054e7 was the lowest compared to both BTC and ETH showing lesser deviation in the average of the predictive dataset’s error calculation. The high MSE values demonstrates the lacklustre quality of the regressed estimator. A high R^2 value of 0.922 demonstrates the model being statistically very significant and a very good fit for the dataset. Similar to BTC and ETH’s dataset, in the case of the one major peaks followed by another big peak that occurring in December 2017, the margin of error as seen in the fitted vs. residuals plot shows heteroscedasticity very similar to that of Bitcoin. This because during the December 2017 bubble period, especially when the transactional volume becomes 10-20 times of almost 90% of the selected dataset.



```
RMSE Squared: 2.4876463095644783E7
MSE: 6.188384161489768E14
MAE: 1.0540201168999761E7
R Squared: 0.9224772965681403
Explained Variance: 5.714027860681208E15
```

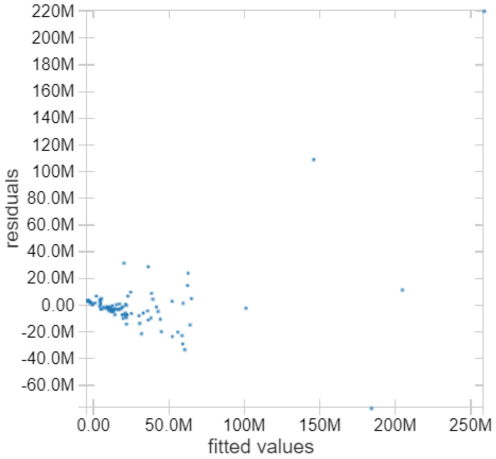


Figure 2.6: Fitted vs. Residuals for LTC Transactional Volume USD

<u>The effect of Opening, Closing and Transactional Volume on Average Pricing of a BTC, ETH and LTC</u>

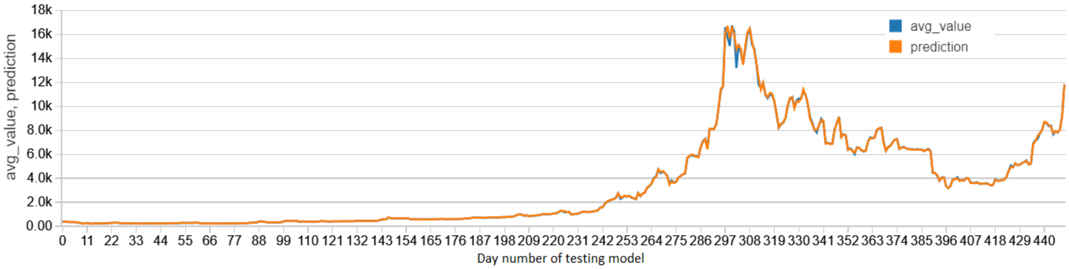


Figure 3.1:Prediction of average value of Bitcoin by using open, close and transactional volume in USD for a given day

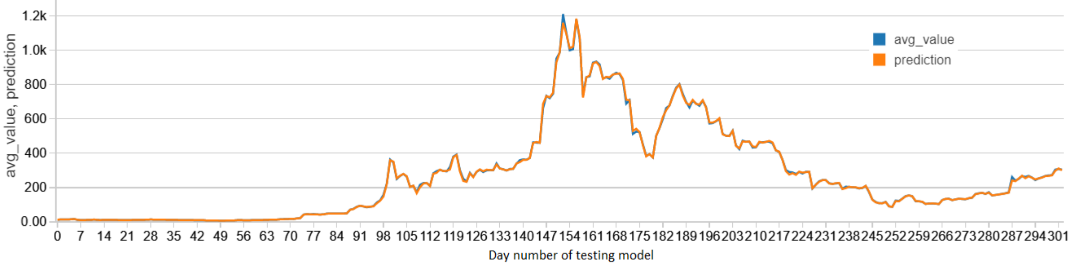


Figure 3.2:Prediction of average value of Ethereum by using the open, close and Tansactional Volume in USD for a given day

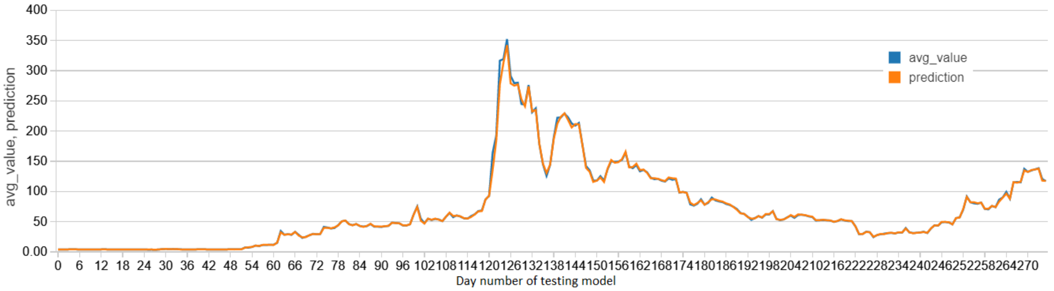


Figure 3.3:Prediction of average value of Litecoin by using open, close and transactional volume in USD for a given day

The following table 3.0 summarizes the regression analytic values:

Table 3.0: Regression analytics for BTC, ETH and LTC

	Bitcoin (BTC)	Ethereum (ETH)	Litecoin (LTC)
RMSE squared	98.350	5.620	3.140
MSE	9672.886	31.591	11.632
MAE	32.404	2.736	1.100
R^2 value	0.999	0.999	0.997
Explained Variance	1.533e7	72343	4249.031

With an R^2 value of 0.99, the average value as determined from open, close and transactional volume in USD yielded statistically great fits for all three cryptocurrency. Despite the model fitting with huge success for Ethereum and Litecoin, the regressed model for Bitcoin generated a relatively high Mean Square error of 9672.886 and subsequently a high Root mean square error of 98.350. MSE being a risk function, corresponding to the expected value of the squared error loss, does not account for accurate predictions but instead measures the quality of an estimator. Therefore, indicating a high risk when utilizing the opening price, closing price and transactional volume USD to determine the average cost of Bitcoin. In contrast to the higher values in Bitcoin, Ethereum and Litecoin have very small MSE and MAE error values showing that the selection criteria for a predictive model for both ETH and LTC is less risky compared to BTC.

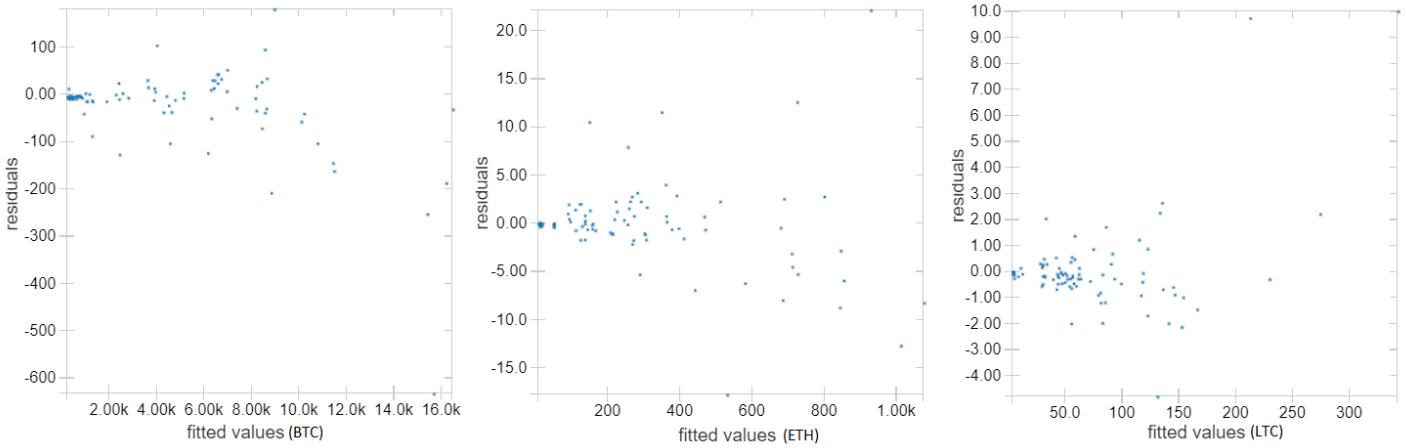


Figure 3.4: Fitted vs Residual plot for Bitcoin(BTC), Ethereum (ETH) and Litecoin (LTC)

The fitted vs. residual plot for Figure 3.4 demonstrates limitations of regressed models for all the datasets. While most fitted values vs. residual generated by Bitcoin would indicate a “no problem” spread, 2 data points, specifically at $16,000 describing the peak values of Bitcoin would demonstrate heteroscedasticity. This can also be seen with Litecoin, where a single data-point at $200 yields a very high residual showing its random variables have increasing variance. Ethereum’s fitted vs. residual on the other hand, has a more random spread indicating a “no problem” fit of the plot. This makes Ethereum’s prediction of its average pricing and its associated variance much more reliable as compared to that of Bitcoin and Litecoin. However, it is to be noted that the model is much more stable for smaller values of ETH as compared to higher values, where there is always a higher associated risk, indicated by a single data-point in the extreme right hand corner of the plot.

The influence of Bitcoin on Ethereum and Litecoin's average price

The opening, closing and average price alongside the transactional volume in USD for Bitcoin were used as key features to predict the average value of Ethereum and Litecoin. Figure 4.1 and 4.3 below show the regressed model as "prediction" plotted alongside the average Ethereum and Litecoin value.

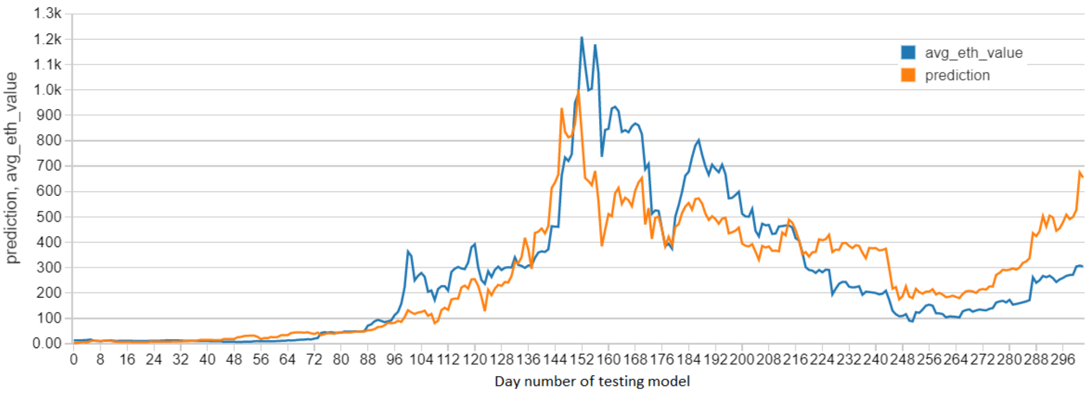


Figure 4.1: Prediction of average cost of Ethereum by using Bitcoin's Transactional Volume in USD, along with open, close and average pricing

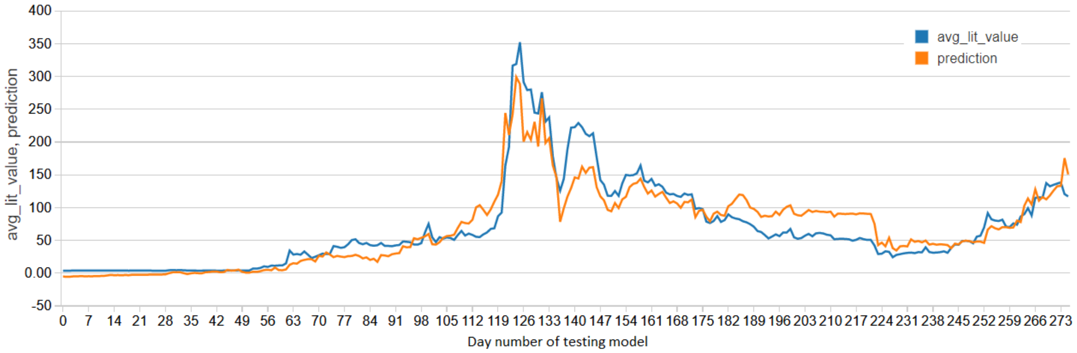


Figure 4.3 : Prediction of average cost of Litecoin by using Bitcoin's Transactional Volume in USD, along with open, close and average pricing

Regression analytics as seen in table 4.0 show very high value for Mean Square error in predicting Ethereum, which being a risk function, makes sense as Ethereum is a higher value currency compared Litecoin and using BTC metrics to predict Ethereum prices can be risky especially at higher evaluations of the cryptocurrency. However, Bitcoin metrics can predict Litecoin average pricing much better due to its lower value lesser comparative Mean square error values and the related RMSE squared value. Both models have a statistically significant R^2 value of >0.7, with Litecoin having a better fitted model with an R^2 value of 0.85. Furthermore, figure 4.2 showing the Fitted vs. residual plot for predicting Ethereum shows more heteroscedastic behaviour compared to that of Litecoin, which has more of a "no problem" spread meaning that random variables do not produce that much of a finite variance.

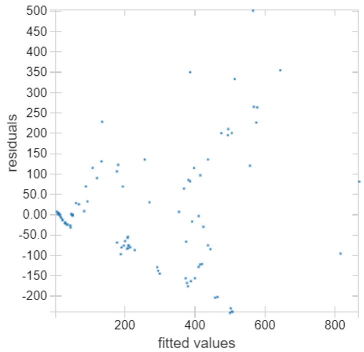


Figure 4.2: Fitted vs residuals plot for BTC predicting

Table 4.0: Regression analytics for prediction of average pricing of ETH and LTC

	Ethereum (ETH)	Litecoin (LTC)
RMSE squared	136.895	25.466
MSE	18740.385	648.537
MAE	96.385	18.456
R^2 value	0.741	0.853
Explained Variance	47285.930	3505.826

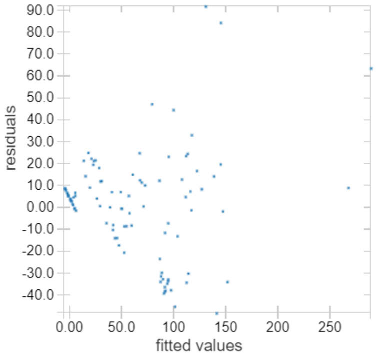


Figure 4.4: Fitted vs residuals plot for BTC predicting LTC average pricing

The influence of Ethereum on Litecoin's average price

The opening, closing and average price alongside the transactional volume in USD for Ethereum were used as key features to predict the average value of Litecoin. Figure 5.1 and 4.3 below show the regressed model as "prediction" plotted alongside the average Litecoin value

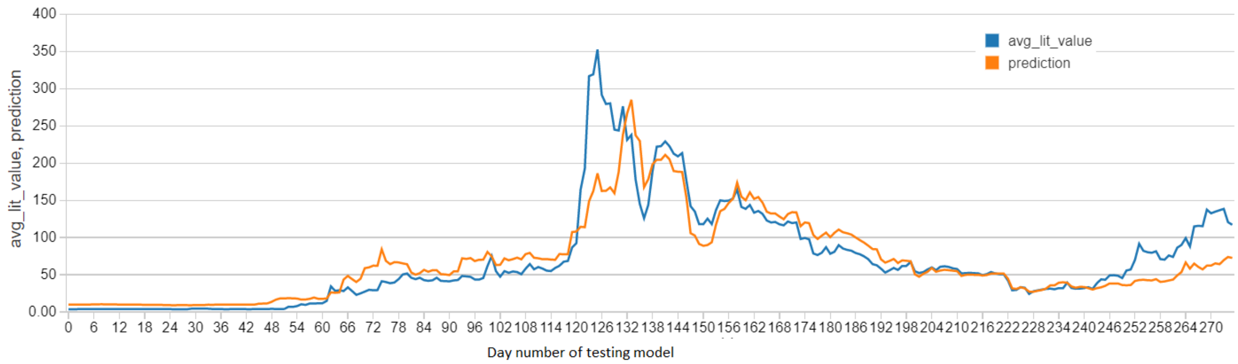
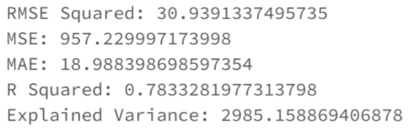


Figure 5.1: Prediction of average cost of Litecoin by using Ethereum's Transactional Volume in USD, along with its open, close and average pricing

Ethereum despite being created after Litecoin, has always had a higher value and transactional volume compared to Litecoin. Figure 5.1 and an R^2 value of 0.78 shows that the regressed model from Ethereum's open, close and average pricing alongside its transactional volume is statistically significant as it more than 0.7 but should not be considered a well-fitted robust model. To indicate this even further, a high Mean square error value of 957.229 would indicate associated risk at higher evaluations of the average price of Litecoin. Despite, the several shortcomings of the model, at lower price points, the ETH's key features can be a good predictor of Litecoin's average pricing. This is further explained by the relatively small explained variance that is detailed in the fitted vs. residual plot as seen in Figure 5.2 that demonstrates slight heteroscedastic behaviour that can be mostly credited to the models poor fitting while predicting the first peak of Litecoin's average price of $200. For lower values of LTC, the model regressed by Ethereum's key features can potentially be a good predictor with less mean absolute error.



```
RMSE Squared: 30.9391337495735
MSE: 957.229997173998
MAE: 18.988398698597354
R Squared: 0.7833281977313798
Explained Variance: 2985.158869406878
```

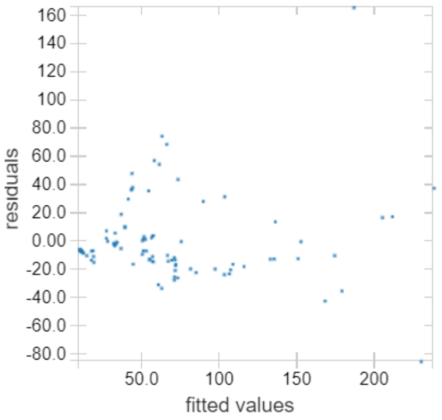


Figure 5.2 : Fitted vs residuals plot for ETH predicting LTC average pricing

The influence of Litecoin on Ethereum's average price

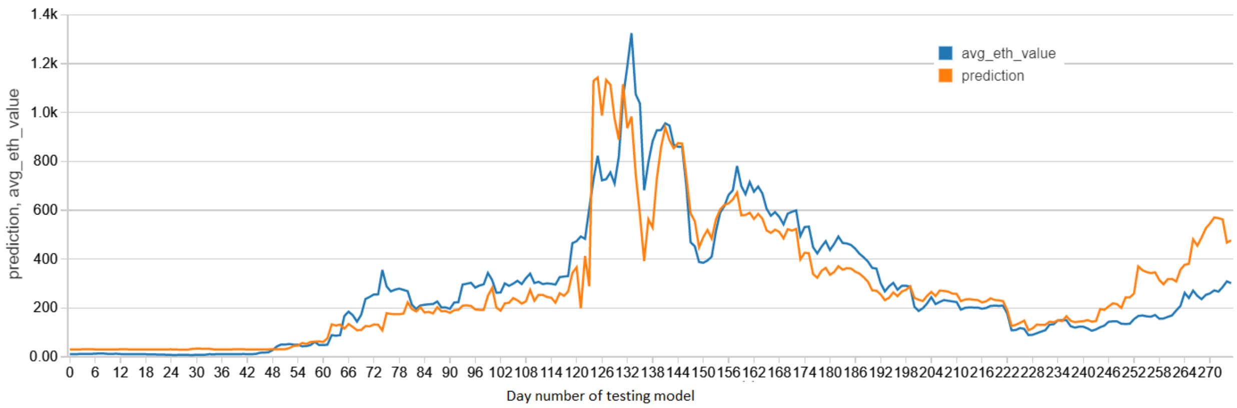
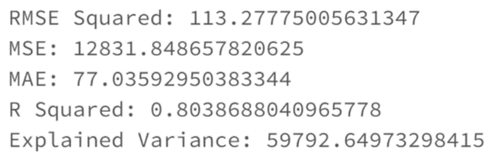


Figure 6.1: Prediction of average cost of Ethereum by using Litecoins transactional volume in USD, wlong with its open, close and average pricing



```
RMSE Squared: 113.27775005631347
MSE: 12831.848657820625
MAE: 77.03592950383344
R Squared: 0.8038688040965778
Explained Variance: 59792.64973298415
```

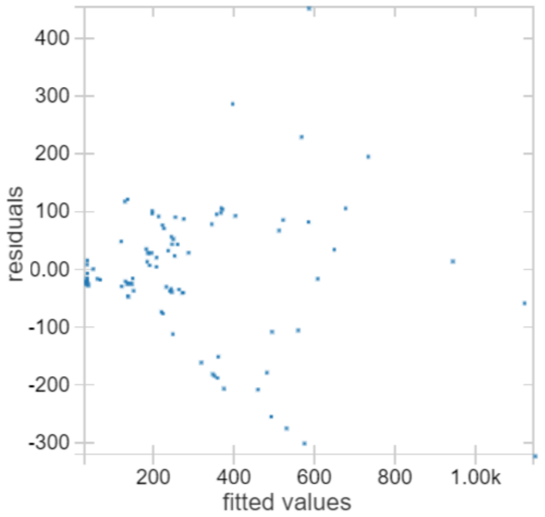


Figure 6.2: Fitted vs Residuals plot for LTC predicting ETH average pricing

The regression model indicates that the influence of Litecoin’s key features of its transactional volume in USD, and including LTC’s opening, closing and average pricing for a given day is weaker when compared to Ethereum’s regressed model. A very high Mean Squared Error of 12831.846 and the subsequent RMSE squared 113.277 indicates high risk when calculating for high price points. As seen in figure 6.1, the regressed model predicts the extent the rise average Ethereum value ahead of time. This is exactly opposite to regressed model in Figure 5.1 where the Ethereum based model under-predicted the rise in cost after the actual first peak. Furthermore, the R^2 value of 0.80 calculated from LTC based model is higher than R^2 that was calculated from model that was regressed using the metrics from the ETH based model. This indicates a comparatively better fit but both models cannot be seen to have great statistical significance. Furthermore, the explained variance of 59792.649 is approximately 30 times higher when compared to the ETH based prediction model. To elaborate, the fitted vs residuals plot as seen in Figure 6.2 demonstrates highly heteroscedastic behaviour showing that the variability of the three given Litecoin metrics are unequal and the variance between the values increases considerably as the fitted values increase in magnitude. Therefore, Litecoin’s transactional volume in USD combined with its opening, closing and average price creates a model that is a weaker predictor of Ethereum’s average pricing, especially when it is compared to the model that predicts Litecoin’s average pricing.

Summary of Conclusions

The results effectively demonstrate how various columns in the three cryptocurrency dataset predict different variable in the same or a different dataset. The high and Low pricing for a given day was used to predict the daily Transactional Volume in USD. The regressed model for BTC and LTC produced great fitting models with a statically very significant R^2 value of more than 0.9. The associated regression metrics produced very high Mean Square Error and Mean absolute error values showing that the regressed model was a very risky model. The most successful 3 models when R^2 values are compared were found to be that for opening, closing and transactional volume in USD of the day being used to determine the average cost for the 3 crypto-currencies. The regressed models were nearly identical to the actual average pricing values. Despite the fit, there was high risk associated with the BTC regression model, but was found to be less risky for ETH and LTC. This was further elaborated by the fitted vs. residuals plot that indicated high variance associated with peaks and extreme change in Bitcoins value.

The influence of Bitcoin’s open, close and average pricing along with its transactional volume produced regression models that were a better fit for Litecoin as compared to Ethereum. The associated regression metrics indicated similar results of high risk associated with high price points and stronger predictions when the values of LTC and ETH are low. Similarly, the Ethereum’s various price points and transactional volume was found to be a better determinant of Litecoin average price than the other way around. All regression metrics indicate the inability of one LTC and ETH to predict peaks in average pricing as the regression models produced significant difference in peak timing of ETH and LTC. As with other regressed models, a sharper change in the cryptocurrency pricing indicated high risk.

Appendix

Due to the large volume of the code and multiple dataset files, the code files (.dbc) and datasets are attached with the submitted document. The files are labeled as mentioned below:

1. How well do the high and low prices determine the transactional dollar volume in USD for BTC, ETH and LTC
    a. Bit hi_lo vs vol_USD.dbc
    b. Eth hi_lo vs vol_USD.dbc
    c. Lit hi_lo vs vol_USD.dbc
2. How well do the opening, closing and transactional volume in USD of the day determine the average cost for the 3 crypto-currencies.
    a. Bit o_c_vol vs avg_val.dbc
    b. Eth o_c_vol vs avg_val.dbc
    c. Lit o_c_vol vs avg_val.dbc
3. How well does Bitcoin open, close and average and transactional volume in USD determine the average cost of Ethereum and Litecoin
    a. Bit vs Eth avg.dbc
    b. Bit vs Lit.dbc
4. How Ethereum open, close and average and transactional volume in USD influences Litecoin average price
    a. Eth vs Lit.dbc
5. How Litecoin open, close and average and transactional volume in USD influences Ethereum and average price.
    a. Lit vs Eth.dbc

The CSV files that attached are named as:

1. Coinbase_BTCUSD_d.csv
2. Coinbase_ETHUSD_d.csv
3. Coinbase_LTCUSD_d.csv