Crypto Currency Price Prediction
Jimmy Nunnally
UMBC
DATA 606 Spring 2020

Introduction:

Crypto currency is a class of notoriously volatile digital assets that is available to trade/invest online similar to other assets like stocks. Due to its extreme volatility, it is extremely difficult to accurately predict where the price is going. At the same time, however, volatility is an opportunity of high levels of profit. My project will develop a crypto currency pricing model. I have picked specific block chain based crypto currency called Ethereum to analyze. My project explores aspects of price, utility, and adoption with the overall objective of predicting the daily closing pricing of Ethereum. The measure of success in this prediction is the ability to profitably trade Ethereum as the end goal is to incorporate my prediction into trading.

There are many possible inputs; one of the difficulties to this model is gathering data from many different sources to feed into the model using a common format. I am pulling the following categories of data from July 2015 to February 2020 for use in the model.

Historic pricing and exchange data

- Open, Close, High, Low, and Volume

Blockchain analysis: (full blockchain is public)

- Unique addresses
- Distribution of assets
- Number of transactions

Network effect:

- Search engine query trends

Economic and financial inputs:

- Economic data
- Fear index
- Stock market/commodity data

Existing approaches:

Due to crypto currency price forecasting being somewhat of a profit driven subject, there are numerous projects out there that attempt to forecast crypto currency. I identified three similar projects that are not locked behind a “paywall”.

The first project is a capstone from Berkley.[1] The details are unavailable, however, the project summary is. I found that they use intuitive inputs that are similar to mine. They are also claimed a high degree of success using the random forest method, so I feel comfortable that the data I am using will also be of high utility. They also focused purely on Bitcoin.

The second project is a research paper titled “The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy.”[5] This project attempted to make a cryptocurrency price forecast from social media data. Again, they used similar data (exchange and social media trends) to generate a model with predictive ability and were able to identify a feedback loop between social media and pricing. They also focused purely on Bitcoin.

The third paper[4] purely draws from exchange data and tried out numerous machine learning models in an attempt to model short term price predictions for various cryptocurrencies offered at the “OKcoin” exchange. This paper shares my objective, but uses less data than what I intend. The authors recommend Extreme Gradient Tree as the best model to use.

The largest takeaway I gathered was that you really only need historical data from an exchange to support a semi-accurate machine learning model. However, the more data you can gather the more accurate the model can become. It is therefore, my goal, to have as much utilizable elements as possible.

Preliminary Exploratory Data Analysis results

Through my preliminary analysis I discovered that Ethereum transaction fees can explain at least part of Ethereum’s price. I also uncovered five features out of my 38 feature dataset that I want to use in my machine learning model. Transaction count, transaction cost, addresses, unique addresses, and Google trend. I also discovered a high correlation between changes in Ethereum price with change in price in other crypto currencies, with the correlation decreasing as the market capitalization of the other crypto decreases.

Crypto Currency Price Prediction
Jimmy Nunnally
UMBC
DATA 606 Spring 2020

Data Preparation:

Using the results from delivery 2, I have identified features that correlate highly with Ethereum price and merged them all into the same dataframe based upon a datetime index. These features are: number of addresses, number of unique addresses, number of transactions, transaction fee, Google trends, price open, price close, price high, price low, and exchange volume. I have also added in valuable economic data including the prices of Gold, Oil, S&P 500, Dollar Price Index, and Vix. Vix is colloquially known as the "fear" index and represents financial volitilty. The dollar price index shows the strength of the U.S. dollar which has a direct impact on the U.S. dollar price of crypto currency. In depth EDA of these features can be seen in delivery 2.

Model Selection:

When conducting a literature review, I found a similar project called "Predicting short-term Bitcoin price fluctuations from buy and sell orders"[4]. This paper constructed numerous models, but found the Extreme Gradient Tree gave the best results. I therefore researched the xgboost python library (extreme gradient boosting). Gradient boosting uses a collection of ensemble methods to make a prediction while using boosting to decrease computational resources. I found that this is applicable to my dataset which is a time series regression problem and extremely resource hungry.

Time Series Prediction:

A time series prediction is one where the order of events is important. For cryptocurrency pricing this is crucial. This adds a challenge as Time Series data cannot be simply is shuffled randomly or ordering would be lost.[3] Therefore instead of using the randomized test/train split included in scikitlearn , I created distinct time windows to separate my test, train, and validation sets (fig 2). Cross validating my data set is going to be a challenge to tackle during delivery 4.

Feature engineering:

I used financial theory to develop 3 new sets of features. First, I added in moving averages. As prices can change quickly, I am concerned that this can throw off my model. Moving averages smooth overs this variance. Second, I calculated the relative strength index (RSI). RSI is a heavily utilized metric in the investment industry. It is a measure for how overbought or oversold an asset is. Please see appendix for the formula. I believe RSI will prove to be a powerful feature because it is so commonly used by investors. I calculated a third technical metric, Moving Average Convergence Divergence (MACD). MACD is a "trend-following momentum indicator that shows the relationship between two moving averages of a security's price. The MACD is calculated by subtracting the 26-period Exponential Moving Average (EMA) from the 12-period EMA. The result of that calculation is the MACD line. A nine-day EMA of the MACD called the "signal line," is then plotted on top of the MACD line, which can function as a trigger for buy and sell signals. Traders should buy the security when the MACD crosses above its signal line and sell - or short - the security when the MACD crosses below the signal line[2]."

With the addition of the above 3 features my dataset is finalized, I am ready to select a machine learning model.

Initial Implementation:

After creating test/time/validation time windows I called the xgboost library to begin training. This library contains the ability to use cross validation to select the best parameters; however, you need to know the starting point. The preliminary run of model produced an underfitted prediction (fig 3).

To improve upon the prediction, my method was to select 3 different sets of "default" parameters and then develop a new set based upon the best set the cross validation selected. Additionally, To further improve my prediction I decided to feed my model more data. I added in economic data that correlates with Ethereum price including: oil, gold, stock indexes, and the dollar price index.

This produced a better fit, but the results were still disappointing. Something was fundamentally wrong with my model.

Final implementation:

The crypto currency market has changed a lot in the past 5 years. Ethereum traded for mere pennies in 2015 and quickly experienced exponential growth to over a thousand dollars before stabilizing in 2018 at several hundred dollars. I realized that given how much the market has changed, the early data might actually be “tricking” my algorithm. I therefore truncated the dataset to 2018 when the price stabilized. Once I input the truncated dataset into my algorithm, I received much more accurate results.

Measuring Success

I initially started out with a traditional measure of accuracy to measure degree of success that my prediction achieved. The measure is called the “means squared error” and my objective was to minimize this. In this regard I have achieved a high degree of success. I reduced my means squared error to 105 across 132 predicted data point and you can see graphically that my prediction is quite close to the truth. This was roughly a 92% improvement from my initial model which had a mse of 1322. However, my objective is not as simple as getting closest to the “true” closing price, it is to profitably trade. I therefore developed my own profitability metric which mimics real trading to see determine the utility of my model.

To determine my profitability I downloaded the prediction and truth values to csv. I pretended to start with 1 Ethereum at $176.01. If the prediction determined the price would go down the next day I called it a “sell” order and if the price was predicted to go up it was a “buy order.” The difference between the actual values was the profit or loss, which I summed up over the predicted time span. My model correctly predicted when to buy or sell 50.4% of the time correctly. This is about what you would expect from random chance, however, it does appear to correctly predict the larger swings in price. I therefore calculated a total profit of $23.96 over a period of 132 days which is a return 13.6%. See fig(5) for a graphical representation.

Conclusion

My model does a good job at predicting future Ethereum pricing, and is moderately successful at producing a profitable trading model. One key addition that would likely build a stronger model is social media data. The best I could gather on my own was Google Trend to represent social media, however is discussed in "The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy," this information is enough by itself to develop a pricing model [5]. Given what my model does well at, however, I think it would be wise to change from predicting daily closing cost to predicting a rolling average. This would be an easy change to make as it would use the same data I already have and I have already used moving averages as inputs to this model. Daily closing price is affected mainly by chance and is difficult to predict whereas the moving average is smoothed out over time and thus more likely to be predictable. Furthermore, longer term investing is more likely to be profitable due to less friction from trading fees and requires less effort with keeping the model up to date.

Crypto Currency Price Prediction
Jimmy Nunnally
UMBC
DATA 606 Spring 2020

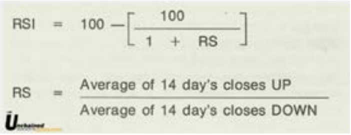
References:

1) “CoinPredictor: One Stop Bitcoin Insights Tool,” *UC Berkeley School of Information*. [Online]. Available: https://www.ischool.berkeley.edu/projects/2018/coinpredictor-one-stop-bitcoin-insights-tool. [Accessed: 1-May-2020].
2) J. Chen, “Relative Strength Index (RSI),” *Investopedia*, 06-May-2020. [Online]. Available: https://www.investopedia.com/terms/r/rsi.asp. [Accessed: 1-May-2020].
3) “Pythonic Cross Validation on Time Series,” *Francesco's Blog* -, 25-Sep-2014. [Online]. Available: http://francescopochetti.com/pythonic-cross-validation-time-series-pandas-scikit-learn/. [Accessed: 1-May-2020].
4) Guo, Tian & Antulov-Fantulin, Nino. (2018). Predicting short-term Bitcoin price fluctuations from buy and sell orders. 25-Sep-2014. [Online]. Available https://www.researchgate.net/publication/323141771_Predicting_short-term_Bitcoin_price_fluctuations_from_buy_and_sell_orders?enrichId=rgreq-c5c3e2551eac9650e731da8639be436f-XXX&enrichSource=Y292ZXJQYWdlOzMyMzE0MTc3MTtBUzo1OTgzMTQwMTM3MDgyODhAMTUxOTY2MDU4NTI4OQ%3D%3D&el=1_x_3&_esc=publicationCoverPdf[Accessed: 11-May-2020]
5) D. Garcia, David Garcia David Garcia, D. Garcia, C. J. Tessone, C. J. Tessone, P. Mavrodiev, N. Perony, N. Perony, Google, Google, Google, Pavlin Mavrodiev Google, and Google, “The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy,” *Journal of The Royal Society Interface*, 06-Oct-2014. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rsif.2014.0623#RSIF20140623F1. [Accessed: 1-May-2020].
6) Wangqiyuan, “Stock trend and prices prediction using XGBoost,” *Kaggle*, 29-Feb-2020. [Online]. Available: https://www.kaggle.com/wangqiyuan/stock-trend-and-prices-prediction-using-xgboost/notebook. [Accessed: 1-May-2020].
7) “Yahoo Finance - Stock Market Live, Quotes, Business & Finance News,” *Yahoo! Finance*. [Online]. Available: https://finance.yahoo.com/. [Accessed: 1-May-2020].
8) *Google Trends*. [Online]. Available: https://trends.google.com/trends/?geo=US. [Accessed: 1-May-2020].
9) *Coinmarketcap*. [Online]. Available: https://www.coinmarketcap.com. [Accessed: 1-May-2020].

Crypto Currency Price Prediction
Jimmy Nunnally
UMBC
DATA 606 Spring 2020

Appendix:

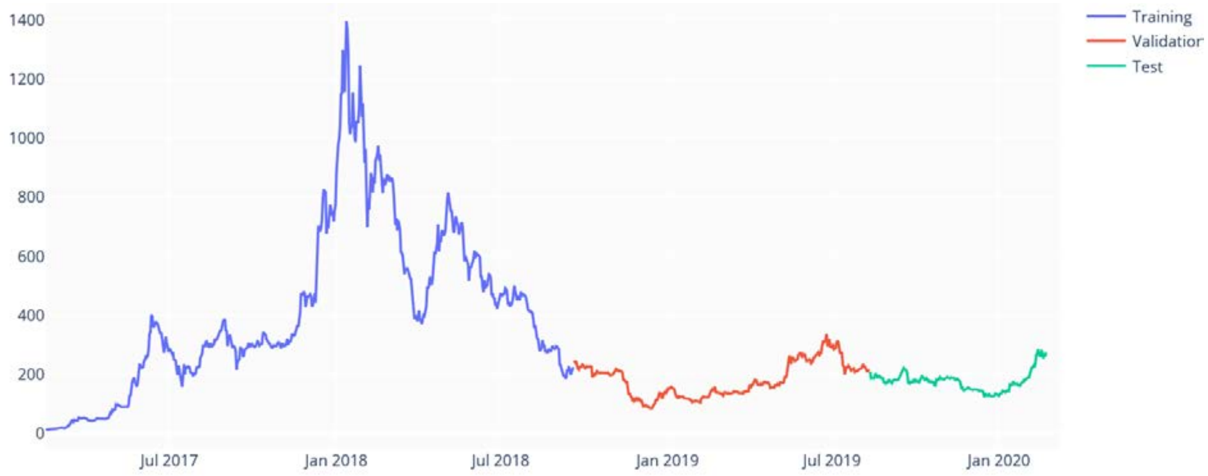
Fig (1)



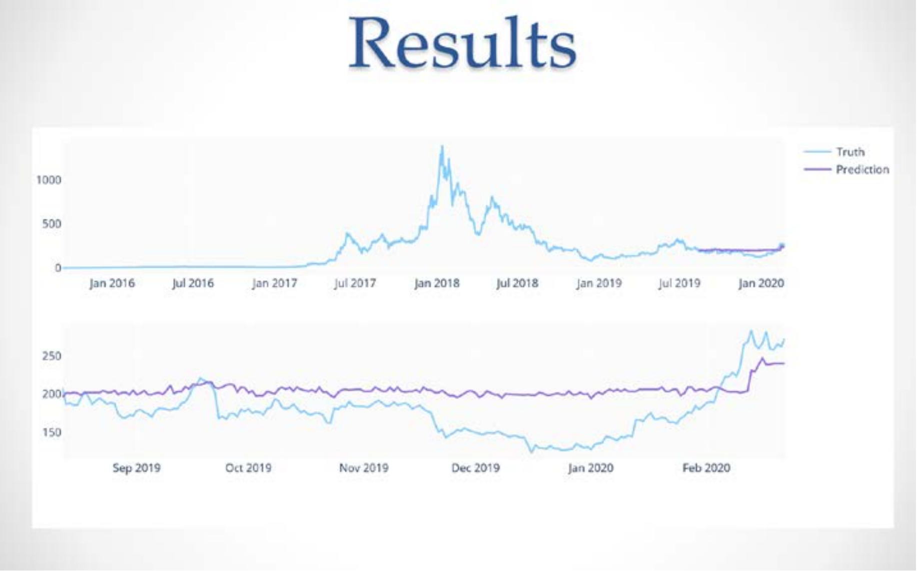
$$RSI = 100 - \left[\frac{100}{1 + RS}\right]$$

$$RS = \frac{\text{Average of 14 day's closes UP}}{\text{Average of 14 day's closes DOWN}}$$

Fig(2)



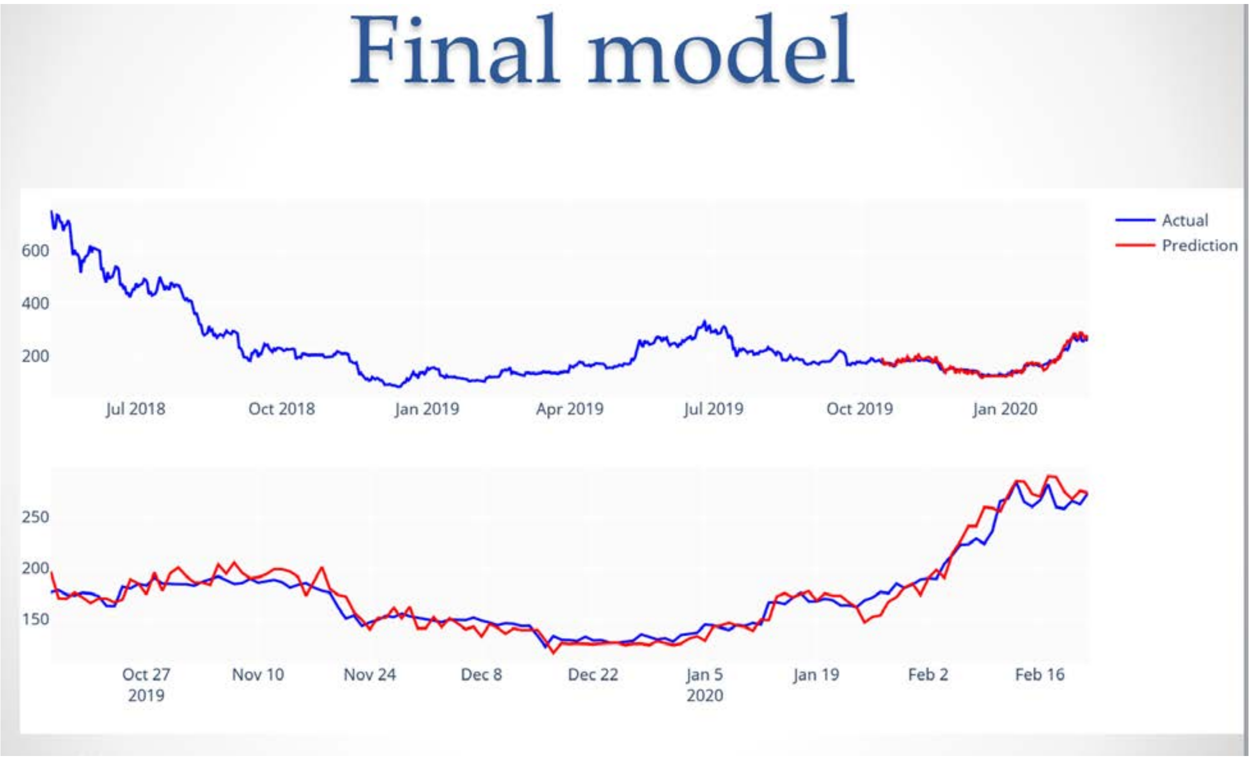
Fig(3)



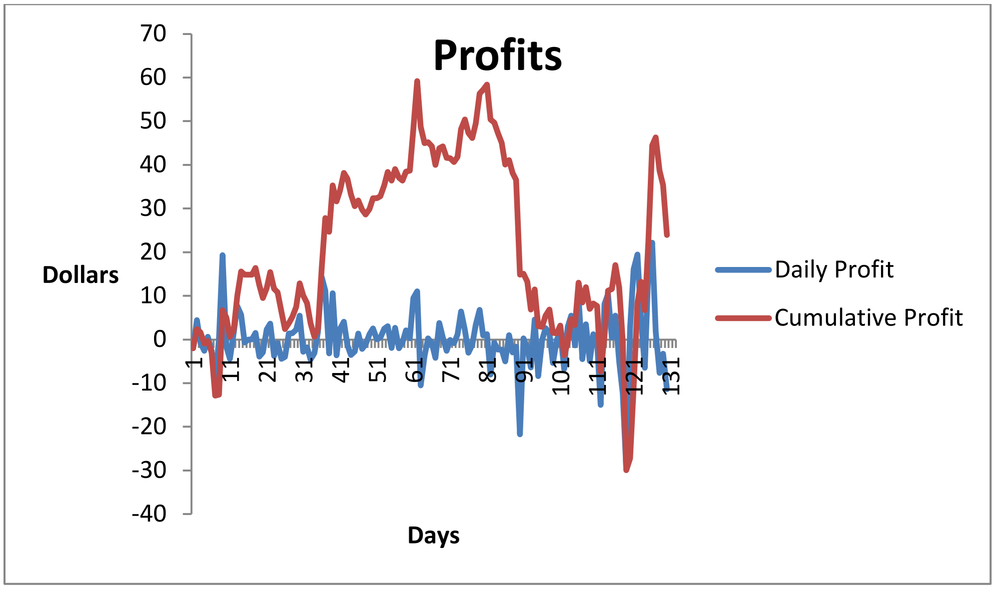
Results

Crypto Currency Price Prediction
Jimmy Nunnally
UMBC
DATA 606 Spring 2020

Fig(4)



Fig(5)


Profits
Dollars
Days
Daily Profit
Cumulative Profit