CS109A Introduction to Data Science:

Milestone 2, example, **adapted** from Ali Dastjerdi et al Spring 2018

Computer Science CS109

Milestone #2 Scope of Work: Cryptocurrency Returns

Project Statement and Background

Cryptocurrencies have divided many in the financial services and technology communities over whether they are more innovation and a boon for society than a fraud. As of April 5, 2018, the market capitalization ("market cap") of cryptocurrencies is over $255 billion. This figure commonly fluctuates by tens of billions of dollars given the immense volatility associated with cryptocurrencies.

Goals: Using the data enumerated below, our goal is twofold: (i) to predict the price return of a cryptocurrency and (ii) using that predictive model, develop a strategy (model) for whether to buy or sell a cryptocurrency on a given day.

We will likely limit the scope of cryptocurrencies we analyze to the following five, which, as of 4/1/2018, were among the top 10 cryptocurrencies in terms of market cap on coinmarketcap.com and have data going back to 2015 or earlier: Bitcoin (btc), Ethereum (eth), Ripple (xrp), Litecoin (ltc), and Stellar (xlm). This allows us to have over 950 data points if we decide to use daily rolling returns in our design matrix and use a time frame from 8/8/2015 (the earliest date all five have data available together) to 3/31/2018. BitCoin and Litecoin have earlier start dates compared to others.

Data:

● Prior daily rolling returns on cryptocurrencies. For example, if we choose to predict the price return of Bitcoin on for the next day (t+1), then we will include the change in the closing price for Bitcoin and other cryptocurrencies on the prior day in our design matrix (the change from t-1 to t). Using rolling returns is commonplace in finance, as it removes the influence that would result from different magnitudes in price levels for cryptocurrencies and other asset classes whose price movement we want to include.

● Price returns on other financial assets (such as equity indices, bond prices, foreign exchange prices, gold, volatility indices of exchanges, prices of major commodities, etc.). It is our hypothesis that the correlation with more traditional financial assets will become stronger (although regarding the direction, we're not sure about) over the time period analyzed as cryptocurrencies become more mainstream in the investment community.

● News pertaining to cryptocurrencies. We will need to create features out of news stories released that relate to cryptocurrencies for incorporation into the design matrix. Major news headlines will also be incorporated as a feature.

What we ultimately hope will differentiate our project from other "studies" on-line includes:

● Incorporation of cryptocurrency-related news. This will require NLP techniques to extract positive/negative signal features.

● A more nuanced treatment of covariance and other financial market data.

Challenges:

● The cumulative returns on cryptocurrencies have exhibited an extremely strong upward trend. For example, the price of Bitcoin rose 2,572% from 8/8/2015 to 3/31/2018. The challenge we will face will be akin to the challenge confronted by Wall Street in building predictive models on technology stocks on price returns (alone) leading up to the tech crisis. It would be difficult for significant exuberance to not be fit by the model.

● Cryptocurrencies exhibit much more extreme variance compared to traditional financial assets, such as stocks or bonds.

● As one of this project's members researched for a project in AM205 (listed in literature below), the covariance between cryptocurrencies has fluctuated significantly. In other words, the comovement of cryptocurrencies is much less predictable than it is, for example, between similar stocks in the oil services sector or for global equity indices.

● Few data points: Using ~950 data points or ~2000 (limiting to Bitcoin & Coinbase) to make inferences may not be enough for applying deep learning methods.

Available Resources/Data

● Coinmarketcap.com: Cryptocurrency data for the top 100 coins by market cap, including: date, open (price), high, low, close, volume, market cap. We have already downloaded,

1 https://github.com/LavenderViking/AM205-crypto-analysis/blob/master/FinalLatex/Project.pdf.

cleaned and saved a local copy as CSV files in our GitHub repo: https://github.com/nate-stein/crypto-ml/tree/master/data. We will also save the data for the other asset classes (once cleaned) there.

● Nasdaq.com: Historical financial market asset data, including: date, open (price), high, low, close, volume.

● FRED: Federal Economic data lists data on economic indices such as bond rates, inflation; foreign exchanges, commodities and currencies

● New York Times Archives: NYT provides an API for accessing the headlines of their historical articles categorized as top stories

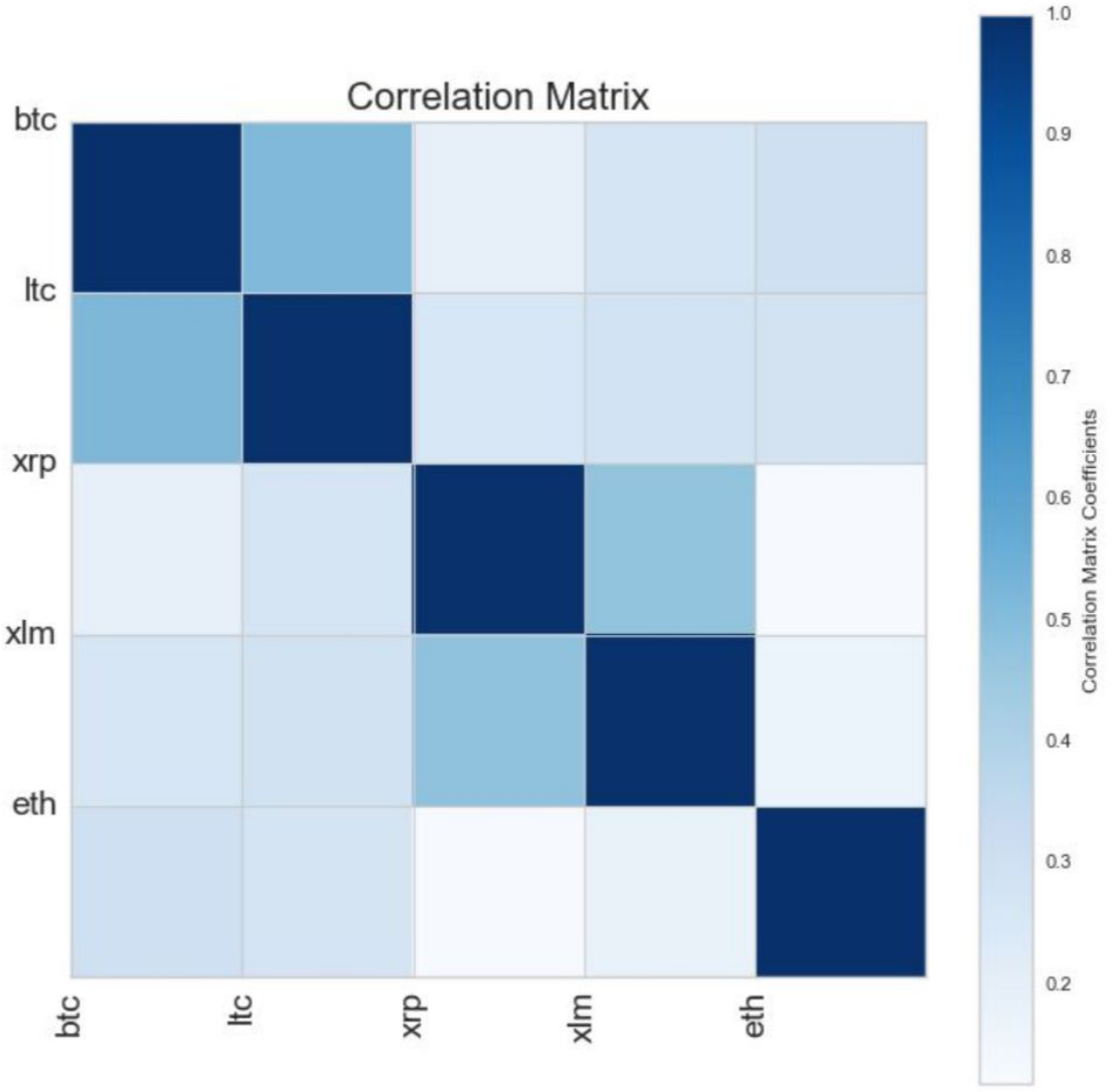
● Crypto Coin News: Crypto Coin news is the primary news source focusing on crypto currencies. This source does not make historical headlines available through an API, and therefore a scraper will be used for data collection.

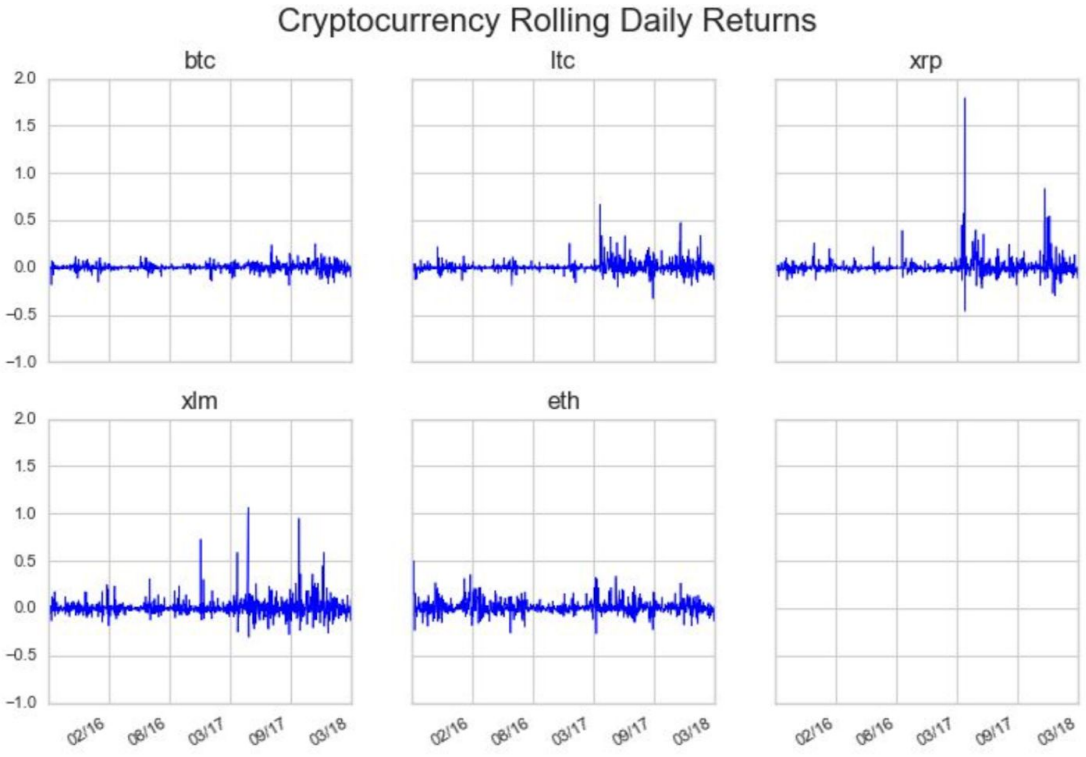
https://github.com/nate-stein/crypto-ml/blob/master/data/file_info.csv lists the data that we have compiled other than the crypto currencies.

Data Analysis Plan:

We plan to remove/impute missing data depending on the variability and number of missing values. Then proceed to check correlations among variables, conduct PCA to check predictors that account for variability in the data. The data would then need to be formatted to be used in the model. We plan to use simpler models such as logistic regression, LDA and proceed to complex models such as random forests, boosting before using deep learning such as RNNs.

Preliminary EDA


Correlation Matrix
btc
ltc
xrp
xlm
eth
btc
ltc
xrp
xlm
eth
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Correlation Matrix Coefficients


Cryptocurrency Rolling Daily Returns
btc
ltc
xrp
xlm
eth
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
02/16
08/16
03/17
09/17
03/18