HARVARD UNIVERSITY
The Graduate School of Arts and Sciences

DATA SCIENCE II

AC109B

Predicting Cryptocurrency Returns

Students:
Ali Dastjerdi
Angelina Massa
Nate Stein
Sachin Mathur

Professors / Advisor:
Pavlos Protopapas
Mark Glickman
David Wihl

May 1, 2018

Table of contents

1 Introduction

Cryptocurrencies are arguably the most polarizing topic within the financial services and financial technology ("fin-tech") community in the last decade. As of April 5, 2018, the market capitalization ("market cap") of cryptocurrencies was over $255 billion, a number that easily fluctuates by tens of billions of dollars given the immense volatility associated with cryptocurrencies. Billions of dollars more are being invested in cryptocurrency-related start-ups. As one would expect, the number of market participants looking to profit from the change in prices of cryptocurrencies has grown immensely.

1.1 Problem Statement and Motivation

Our goal is to build a machine learning model that will enable us to predict the change in the price of a given cryptocurrency that will be realized tomorrow given what we know today. As cryptocurrencies become increasingly accepted as financial assets by mainstream investors, the results from this project and similar predictive modeling exercises have significant implications for investors and market-makers. To build this predictive model, we will experiment with many different types of data, including the trading volume and price history of cryptocurrencies, the price histories of other financial assets, and market-signal features we will engineer using natural language processing (NLP) techniques on financial markets news.

1.2 Data

We focus on the following five cryptocurrencies, 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 enables us to have over 900 data points (using 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 relatively earlier start dates compared to others, but we wanted to encompass a broader universe of cryptocurrencies than just two. Connecting this back to our problem statement, our goal is to build a model that would enable us to predict the price return of one of these five cryptocurrencies by using information on that cryptocurrency's own price history, as well as that of the other four cryptocurrencies and other features.

In addition to using information about the price and volume history of cryptocurrencies, we will also experiment with the 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 one of our hypotheses that the correlation with more traditional financial assets will become stronger (although regarding the direction, were not sure about) over the time period analyzed as cryptocurrencies become more mainstream in the investment community.

1.2.1 Modeling Time Series

Before delving into the existing literature around this topic, it will help to provide a quick overview on certain features of the problem that will accelerate the reader's understanding of following sections. First, as is common in nearly all financial time series, we are concerned with predicting price *changes*, not the absolute price levels. Therefore, our feature space will consist of relative *changes* in the price and volume traded for certain assets. Moreover, because we are more interested in the trading perspective rather than investing, most of our feature space will consist of *daily* moving windows, as opposed to weekly or monthly, which we discuss in more detail later. A brief visual may help; the following figure shows what the rough work-flow would be if we were attempting to predict the price change for Bitcoin the following day based on the price changes for Bitcoin, Ethereum and Ripple:

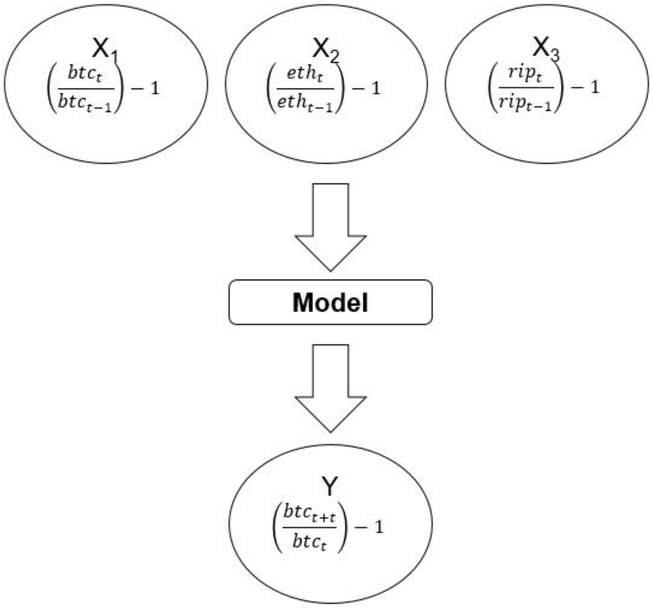
X_1
$\left(\frac{btc_t}{btc_{t-1}}\right) - 1$
X_2
$\left(\frac{eth_t}{eth_{t-1}}\right) - 1$
X_3
$\left(\frac{rip_t}{rip_{t-1}}\right) - 1$
Model
Y
$\left(\frac{btc_{t+t}}{btc_t}\right) - 1$

Figure 1: Simple Model Illustration

As described later, the actual feature matrix we use will be significantly more involved. One caveat is that trade-by-trade data may have been more appropriate for attempting to model cryptocurrency price changes from the trading perspective. However, we were unable to find readily available trade-by-trade data for multiple cryptocurrencies, and so we decided to focus on the closest alternative: day-by-day data.

2 Literature Review

Unfortunately, there is a dearth of sophisticated analysis on cryptocurrency price return forecasting. Although articles with titles such as "Use Machine Learning to Predict Bitcoin" abound, they are geared more towards an audience interested in learning software engineering / data science by applying a smattering of techniques to an oversimplified version of the problem. To remedy these inadequacies, we aim to contextualize the problem of cryptocurrency price return forecasting within the broader literature on financial time series analysis.

Investment approaches are generally bifurcated between (i) "fundamental" investors, those who value companies based on some estimation of the company's "true" value based on its assets, profitability, cash flows, etc., and the (ii) "quantitative" ("quant" for short) investors, who use statistical patterns and machine learning techniques to predict future price changes without regard for the "true" value of their investments. With cryptocurrencies, no attempt has really been made to estimate their "true" value, which is not surprising given how relatively new the underlying technology is and to the fact that cryptocurrencies do not lend themselves easily to any sort of parity valuations used with traditional currencies (such as the EUR/USD exchange rate). We focused on "quant" research for stocks and other traditional financial assets with the aim of drawing parallels to the cryptocurrency focus of this paper. The ultimate hope for these endeavors was that they would inspire ideas for feature engineering and other approaches to our problem. Each of the following subsections centers around summarizing a particular quantitative approach and noting how the existing research influenced how we explored our problem.

2.1 Momentum

Momentum is the "tendency of assets with good (bad) recent performance to continue overperforming (underperforming) in the near future" (Vayanos et al 2013). Han, Yufeng, et al (2013) note that this has been one of the most robust empirical tendencies in financial markets, even though it is, in a way, counterintuitive to those subscribing to the "buy low, sell high" mentality. The most common "technical analysis" strategy to profit off this effect is by using a "moving-average" (MA) strategy whereby an investor buys or continues to hold an investment in an asset when the prior day's price is above its 10-day MA price or invests it into a risk-free asset (such as the 30-day Treasury bill) otherwise. To allow for this effect to potentially be captured by our model, we chose to engineer features capturing whether a given cryptocurrency is above its x-day moving price, where we experimented with several different x-values, i.e., moving-average windows.

2.2 Trading Volume

Vayanos et al (2013) attribute the cause of the momentum effect to some "shock" (e.g., a news event) that impacts the fundamental value of some assets. For example, in February 2018 a Forbes story broke out that major banks, including J.P. Morgan Chase, Bank of America, and Citigroup would no longer allow their customers to purchase cryptocurrencies using their credit cards. This news could be considered a negative shock to the supposed utility of cryptocurrencies as transactional devices and stores of value. Vayanos et al (2013) argue that the momentum effect occurs as a result of the *gradual* outflows that proceed such a news event, preceded by the large sell-off by major investors. Brown et al (2009) further suggest that trading volume could indeed be a proxy for a number of other important factors, such as liquidity, momentum, and other information. They find that "past trading volume predicts both the magnitude and the persistence of price momentum." Therefore, trading volume data for cryptocurrencies may enable us to capture momentum and other effects.

2.3 Covariance between Cryptocurrencies

Any trading strategy relying on historical price patterns of multiple assets should be especially concerned with conducting a nuanced analysis of the covariance between those assets. This is potentially the most lacking area in existing literature on cryptocurrency price returns. Yan et al (2017) explore the covariance between cryptocurrencies from a variety of different angles and conclude that the covariance between cryptocurrencies, although high, is particularly unstable. This relative instability is true whether we use a basket of stocks from the same sector (such as a basket of energy stocks) or a basket of large-capitalization (large-cap) stocks as our point of reference. This is important for our problem in many ways. First, any incorporation of covariance (even if implicit by virtue of including multiple cryptocurrencies in the design matrix) should permit rolling windows to adjust in some form to "learn" any changing comovement patterns. Second, this instability will make it difficult to capture any serial dependence in cryptocurrency returns. DeMiguel et al (2014) use vector autoregressive models to select between equity portfolios and find that the expected returns are greater when the principal components can be used to determine which stocks will experience a positive (negative) return in the future. At its least promising, the work done by Yan et al (2017) suggests this would not be possible with cryptocurrencies, while at the most promising, urges caution when basing a model on any assumption of stable comovement.

The underlying belief in many variants of the autoregressive models is that a given asset occasionally must "catch up" to other similar assets. As an example commonly given for a traditional attempt at arbitrage, suppose the stock price of PepsiCo, Inc. (ticker: PEP) rises significantly today for no readily identifiable reason, i.e., no news or other announcement was released that would provide new information on the health of PEP as a company. If a very similar stock, such as The Coca-Cola Co (ticker: KO), saw no significant rise in its price today, the argument goes that KO's price will catch up the following day, or, rather potentially, in today's high-speed-trading environment, within hours or seconds.

This summarizes the description of where we drew inspiration for some of the kinds of features we engineered. The following section will review how we went out about framing our modeling problem and what our process was in determining which features to maintain, tweak, or remove entirely.

3 Modeling Approach

Our general approach to modeling mirrors the "Machine Learning Development Framework," which is illustrated in the following graphic borrowed from a presentation titled "A Framework for Applying Machine Learning to Systematic Trading":

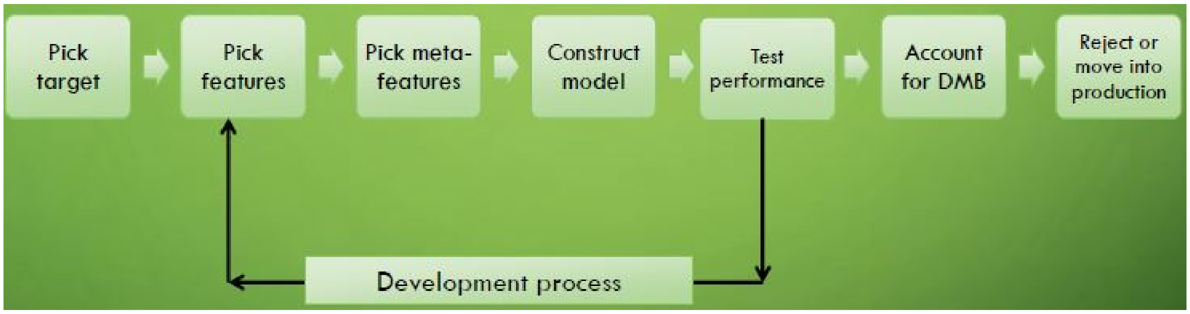

Figure 2: Machine Learning Framework, taken from Kris Longmore of Quantify Partners

Therefore, the organization of this section will largely mirror the above framework.

3.1 Selecting Target

The objective of this project could be addressed in two ways by attempting to predict two related targets:

1. The *actual* price return of a given cryptocurrency, which is a regression problem.

2. The *direction* of the price return for a given cryptocurrency, which can be viewed as a classification problem, with outputs corresponding to trading signals: (+1) Buy (-1) Sell (0) Do Nothing.

3.1.1 Measuring Performance

The selection of a target also involves the selection of how to measure performance. In the regression problem, the most easily interpretable performance metric is the mean absolute error (MAE), i.e., the average absolute difference between our predicted price return and the actual price return. The more common mean squared error (MSE) metric is less intuitive since we are typically dealing with decimal numbers smaller than one, and consequently whose square is actually a smaller number. In the classification case, we could approach the problem in a number of ways. We'll compare performance across two metrics: (i) the percentage of Buy/Sell signals that are directionally correct and (ii) the average returns of executing on the Buy/Sell signals. One additional decision to make with the classification problem is what threshold to use when labeling our training data a 'Buy'/'Sell' as opposed to a 'Do Nothing'. Presumably, we would want to avoid trading when the Buy/Sell signal isn't very strong, so we experiment with varying thresholds.

3.2 Feature Engineering

This section provides an overview of the feature space we experimented with. The following section will describe our process for choosing which features to keep in our model.

3.2.1 Rolling Price and Volume Changes

Given the methodology suggested by the existing literature in systematic trading, we first introduced lagging variables for the price returns of cryptocurrencies. Implicit in the inclusion of these is the belief that future returns are, to some degree, a function of prior returns. First, let us frame a specific problem to make it clearer. Let's say that on March 28, 2018, we wanted to predict the change in the price of Bitcoin that would occur on the following day (March 29, 2018). We further assume that the decision-maker would wait until the end of the day to make a decision and hence would have at their disposal the return experienced by the cryptocurrencies in scope at the end of March 28, 2018. Therefore, we define the trailing ("rolling") returns as:

$$r_t = \frac{\text{price}_t}{\text{price}_{t-1}} - 1$$

where price_t is the closing price on 3/28 and price_{t-1} is the closing price on 3/27. As an example, Bitcoin's closing levels on 3/28 and 3/27 were 7954.48 and 7833.04, respectively, so its daily rolling return on 3/28 is equal to:

$$r_t = \frac{7954.48}{7833.04} - 1 = 1.55\%$$

We created similar lagging features for the volume traded in cryptocurrencies (where volume is defined as the units bought/sold multiplied by the average price at which those units were bought/-sold).

3.2.2 Preprocessing

When using lagging financial time series as features, one must make a decision on how to standardize them. A 10% one-day change in the price of Bitcoin, for example, relays different information depending on when that change was observed. In Figure 3, we see that the standard deviation of daily returns has changed significantly over time. When we compute the standard deviation of daily returns over rolling 30-day windows, i.e., taking the standard deviation of the daily returns in the 30 observations leading up to and including t, we see σ is significantly greater towards 2018 than it is in 2016.

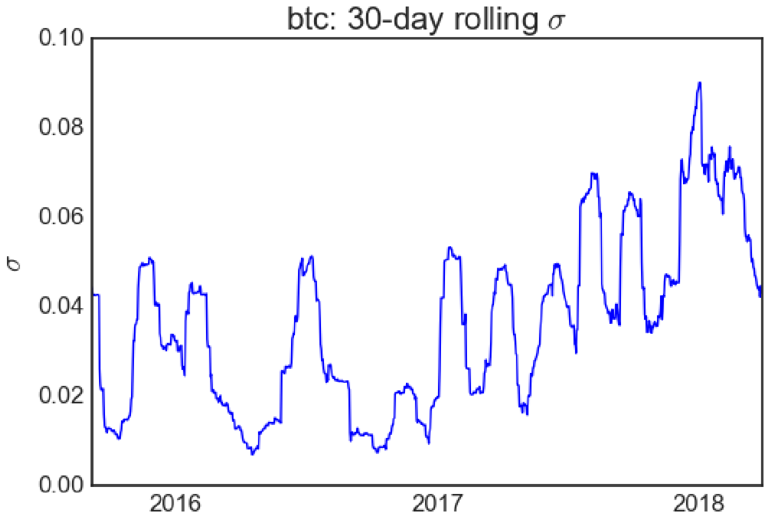

Figure 3: σ calculated over rolling 30-day periods

Returning to the original example being addressed, a 10% one-day change in the price of Bitcoin is more than a 2-σ event for most of 2016 while in 2018 it begins to approach being only a 1-σ event. Therefore, we decided to standardize each daily rolling return as follows:

$$r_t^* = \frac{r_t - \mu_{t-1}}{\sigma_{t-1}}$$

where

$$\mu_{t-1} = \frac{1}{30} \sum_{i=t-31}^{t-1} r_i$$

Of course, the number of days to use for this rolling standardization window is, in itself, a meta-feature to our problem.

3.3 Baseline Model

We first made three decisions regarding the feature space that greatly simplified subsequent steps of further engineering features and deciding which to drop:

1. Fix the rolling window for measuring changes in price and volume.
2. Fix the rolling window used to standardize price and volume.
3. Decide whether to include weekends or not.

First, by deciding on the above factors, we preclude the need to iterate over all different variations when experimenting with different models. Second, our more engineered features (with their own meta-features) depend on how the above steps are set. Finally, it makes intuitive sense that these decisions are relatively model-agnostic. In other words, if a simple model performs better given a certain rolling window for changes in price and volume, then there is strong justification for believing that features created using that window carry the most relevant information, even with more complex models.

After trying out varying rolling windows for

3.3.1 Linear Regression: Continuous Target

The baseline model for the continuous target objective was ordinary least squares regression. It serves as a great baseline model because it is the simplest in terms of interpretation:

3.3.2 Logistic Regression: Categorical Target

4 Results

4.1 Project Trajectory

We were surprised with several developments along the way of our project. The most surprising was the fact that the incorporation of rolling returns for other financial assets did not improve our model's capabilities (for either regression or classification). In hindsight, we should perhaps not have been surprised given the correlation between the daily returns on Bitcoin (as an example) and other financial assets did not exceed 10%. Cryptocurrencies are very different than other assets, such as equities, bonds, currencies and (physical) commodities.
For the baseline model, we first had to decide which rolling windows to use.

4.2 Interpretation

What can we learn from the ultimate choice of models and features? First, the best prior data to use in predicting tomorrow's returns are today's data (and no more than that). Increasing the window with which we calculated the rolling returns up to the present by more than one day for inclusion in our feature matrix worsened the results. Our prior conviction was mixed. Calculating the rolling returns over a longer window could plausibly be viewed as an indirect way of preventing overfitting by capturing a more systemic trend than captured in a single day. Yet, the results proved the reverse. Perhaps this makes more intuitive sense once we recall the immense volatility associated with cryptocurrencies and the fact that

Although some of the results were disappointing, the information we can derive from those results is still immensely valuable. The futility of including returns data from non-cryptocurrency assets in our predictive efforts suggests that the investor base for cryptocurrencies is still very different than that of more traditional financial assets. One way the investment community could interpret this result is that cryptocurrencies offer a form of diversification perhaps unparalleled by other assets.

In the capital asset pricing model (CAPM), investors should be happy with lower expected returns for assets that have a low β (correlation to the overall market) due to the diversification benefit such assets provide, realized through their effect of lowering overall portfolio variance. Although the variance of cryptocurrencies individually is quite high, we would still expect to see a drop in overall portfolio variance, assuming the portfolio were well diversified, because of the seeming lack of relationship between cryptocurrency price returns and the returns of other financial assets.

5 Conclusion

The aim of this paper was to build a predictive model for future cryptocurrency returns based on current cryptocurrency returns and other data. We succeeded in building a model that can predict better than chance alone (at least, when utilized on the time frame covered by this project), but perhaps not one we would feel comfortable betting our own capital on.

5.1 Possible Future Work

Much to do.

References

Brown, Jeffrey H., et al. “Trading Volume and Stock Investments.” *Financial Analysts Journal*, vol. 65, no. 2, 2009, pp. 6784.

Cheng, Evelyn. “JPMorgan Chase, Bank of America & Citi bar people from buying bitcoin with a credit card.” *cnbc.com*. 2 Feb 2018.

DeMiguel, Victor, et al. “Stock Return Serial Dependence and Out-of-Sample Portfolio Performance.” *The Review of Financial Studies*, vol. 27, no. 4, 2014, pp. 10311073.

Han, Yufeng, et al. “A New Anomaly: The Cross-Sectional Profitability of Technical Analysis.” *The Journal of Financial and Quantitative Analysis*, vol. 48, no. 5, 2013, pp. 14331461.

Longmore, Kris. “A Framework for Applying Machine Learning to Systematic Trading.” *Quantify Partners*. RobotWealth.com.

Vayanos, Dimitri, and Paul Woolley. “An Institutional Theory of Momentum and Reversal.” *The Review of Financial Studies*, vol. 26, no. 5, 2013, pp. 10871145.

White, Halbert. “A Reality Check for Data Snooping.” *Econometrica*, vol. 68, no. 5, 2000, pp. 10971126.

Yan, Yihang, et al. “Do cryptocurrencies move in a parallel manner?.” *Harvard University Project*, 2017.