Capstone Project 1: Milestone Report

Problem Statement:

I will be trying to create a model that takes in past price and volume trading values of various cryptocurrencies to predict short-term future prices. One challenge here is to tackle a dataset that inherently contains dependencies with a multitude of real-life events and technological changes that cannot be represented in the dataset. Another is the need to keep a longer memory of the current state than simply current price. In this case, we will have to consider a certain range of past values incorporated into memory. The reason I'm choosing cryptocurrencies is because it's still quite a volatile and immature market that offers volatility that allows for short-term trading.

There are also other indicators that we could possibly extract to provide more signal to our model. One is the MACD, a measure of positive or negative momentum that is conventionally used in such applications. Current returns could also be another indicator of future returns. If you assume a certain volatility in returns then a high current return could result in a reversion to the mean, meaning we should be seeing a negative future return.

This is all complicated by the overall macro-trend of these assets. In the period we are considering, there were very pronounced long-term movements. As a result, our methodology should be assessed against a simpler strategy of a single long or short across the duration of the period assessed.

This exploration will provide insight into the accessibility of profitable technical trading to individual traders. It's a well known fact that large investment firms and hedge funds invest a huge amount of money to gain any sort of advantage for their algorithm, such as laying fiber optic cables to the market so they have access to new information fractions of a second before other firms, investing in massive computing power or higher the brightest minds from a multitude of natural science and statistical fields to create hugely complex black-box models. These models become so far removed in their inner workings from the fundamentals of the underlying asset that they become entirely incomprehensible and inaccessible to us humans. One infamous story is that of a correlation Renaissance Technologies between the weather in Paris and the subsequent performance of the

markets. It was only true slightly more than 50% of the time, but if you aggregate a multitude of such signals, there is an opportunity to achieve high profitability.

Finally, trades cannot be assessed simply by reward, but also by risk. Thus, after the model provides us with a choice of various trades, we will then filter these trades by some metric of riskiness.

Dataset:

My data set was acquired from an API connecting to Cryptowatch, a popular charting and trading terminal that aggregates prices across various exchanges. I chose to use their Bitfinex values, as Bitfinex is one of most used exchanges in the space, as ranked by volume. It was important to extract all values from the same exchange, as there is often a substantial discrepancy in prices across exchanges due to the differences in liquidity and regulatory environments. For example, there was a time when cryptocurrencies were substantially cheaper in China than in other countries due to a ban on fiat-denominated exchanges that vastly decreased the liquidity in the nation.

The period of analysis was from the 1st of January 2018 to the **27th June of 2019**. The interval between data points is 4hrs, allowing for a more discrete analysis of short-term movements. The actual cryptocurrencies were chosen according to percentage of market share, as smaller-cap currencies tend to experience much more volatile trajectories as single transactions tend to cause large immediate changes in price that are harder to predict. We ultimately chose the following currencies: Bitcoin (BTC), Litecoin (LTC), Ethereum (ETH), Ripple (XRP), Bitcoin Cash (BCH), eos (EOS), Bitcoin SV (BSV), Tron (TRX), Stellar Lumens (XLM).

I first extracted the values into a 2D numpy array (price and volume) for each coin, amassing the whole dataset into a dictionary linking currency to the corresponding array. I then built a pandas DataFrame that allowed for a cleaner inspection of the data.

To prepare the dataset, I calculated the log return of the price. This allows for a better distribution of values for coefficient calculation. I also normalized the values of volume. To fill the missing values, I used a 3rd spline fill. Some of the currencies chosen were dropped because they contained too many missing values at the start that the fill gave very strange results.

Exploratory Data Analysis:

In this chapter I put forth several questions which some statistical analysis will try to answer:

1. Is there any correlation between different currencies? If so, is there a delay between this correlation that could be used for prediction purposes?

2. Is current volume a useful indicator of future price changes. Is it simply the absolute volume or the percentage change from one period to the next?

3. Is the MACD a useful indicator of short-term future prices. MACD captures changes in momentum in a time series, but can this indicator be used for short-term trading?

Several statistical methods were used to answer the above questions. The focus here was to find if a correlation with a lag could be found. The point of that is a lag could be useful for prediction. For example, if an increase in volume precedes an increase in prices, volume can then be used as a feature in our model.

For volume, the maximum signal you have is when predicting one period in the future. Beyond that, there is very little correlation between volume (or volume changes) and future periodical returns and total returns over that period.

Regarding prices in different currencies, the maximum correlation was found at zero lag, at 0.78. At any type of lag, this is reduced to negligible values. Interpreting this, we find that while cryptocurrencies prices are highly parallel to each other, it is , at our level of discretization, synchronized and as such cannot be used to predict future prices.

I then tested the MACD indicator across various spans on various points in the future. None of the different MACD spans I tested were able to correlate well with the immediate short-term, but they did have value over a slightly longer period. For example, the difference of the 4-period exponentially weighted moving average from the 2-period exponentially weighted moving average has a 0.04-0.06 correlation with returns over 3-4 periods respectively.