Big Data Applications: Stock and Crypto Trading

Trading has vastly evolved over the past couple decades. First with the advent of computers for faster computation. Then with the creation of the internet for instant access to data. And finally, with the fascination with big data analytics and how to squeeze data points and correlations out of everything. Information after all is the key to market success as theorized and proven throughout the years. While this was initially based on raw financial data, the advent of mass media and social media are now widening the sphere of useful economic data. With more information in all forms comes a more accurate picture of the market landscape that can be pieced together by big data analytics.

The market and its profitability have always been driven by the effective utilization of information. In the early 20th century that meant the analysis of reports and information months old, taken down by hand, with much of the information going unrecorded. Analysis was performed on information that was produced in a slow but steady drip. With the advent of the internet, that drip is now a flood of data financial analysts find themselves wading through. This information gives us a more complete picture of the market, though there is too much to process all at once. Thus, models are created, tweaked, and interpreted to find small yet meaningful correlations in this torrent of data.

For some time, most scholars thought that the market was unpredictable. It was influenced by too many factors to reasonably predict. News, politics, financial analysis, greed, and fear, these are all things that drive each individual investor’s strategies. However, the market is a culmination of all these people whose actions influence each other in subtle ways. This is what is known as the Effect of the Sheep Flock; market prices and value are ultimately driven by the collective interests of all investors involved.

Eugene Fama recognized this and formulated the Efficient Market Hypothesis. He proposed this theory in the 1965 as information infrastructure around the world was taking off. It puts forth that all market participants are rational and pursing maximum economic gain. In efficient markets, markets whose participants have access to as much market data as possible, prices reflect all information available. Thus, all stock prices quickly respond to new information. This puts data accessibility at the forefront of profitability within the market. He classified data availability into three separate forms.

“

Weak Efficiency: Only past information is considered.

Semi-strong form: All publicly available data is used.

Strong Form: All information publicly and privately available is used.

“

His theory has been proven through the years and has been a valuable foundation for forming models of the market. With the advent of Big Data analytics and computational tools, financial analysts are finally able to process copious amounts of data that are not humanly possible. Our capacity to utilize and interpret this data is growing ever closer to the ideal of his semi-strong form. This has completely changed the trading game, as a former senior financial managing director put it “Trading is no longer a balls job, it's a brains job”.

The data that goes into creating an accurate market model today are split into two separate categories: market data and economic agents. Market data is relatively simple, it is the price data, macroeconomic data, fundamental data, etc. Any numbers relating to stocks and analysis of stocks is considered market data. Economic agents are all data we can get from the investors themselves. Before this was primarily news sites and media organizations that would publish ‘sentiments’ on stocks and companies. Whether or not the sentiments are news or opinion, they would be read by investors and weigh on their own personal investment strategies. This can be explained by both the Effect of the Sheep Flock and Efficient Market Theory. When speculation and doubt is cast into the viability of a stock by a media organization that has a high reach in the economic community, most them will react as rational individuals to the new information and possibly sell. This devalues the stock which causes more to be cast into doubt and sell. As this trend continues most of the flock moves away from the stock. This is an exaggerated example, but there are plenty of cases that share these plot points. In as early as 2010, two researchers from Turkey developed a program that would process news articles and predict if stock prices for the companies would rise or fall based on the vocabulary used in the article.

“We accomplished the stock prediction system using financial news articles. Our system automatically analyzes and classifies news articles and generates recommendations for investors. We acquired 61% accuracy from our study.”

A 61% prediction rate is amazing considering you only need 51% to turn a profit. However, it’s not just news articles that are relevant anymore, but social media.

Twitter and Facebooks are hives of information about peoples’ personal opinions and investment strategies. While some of these individuals have less reach than new sites, some of these individuals may have more. Either way there is a constant production of sentiment and opinion that may affect an investors markets strategy and thus the market itself. Information tends to replicate, and sentiment can go viral having a large effect on the market. Another study conducted by a group of researchers from India performed a study on social media sentiments input on the market landscape in conjunction with historical economic data.

“[Our] Model shows that sentiment analysis of the social data complements proven technical analysis methods such as regression analysis” … “Exploiting social media data in addition to numeric data increases the quality of the input and gives improved predictions. The aide of big data technology allows predictions at real-time”

There is a clear benefit to maximizing the amount of data you consider when you try to produce a market model. Whether it be stock data or data on the sentiment of investors themselves, all the data gives you a more complete picture of the market landscape. For my research topic I wish to exploit this to create a market model utilizing big data that will hopefully have a prediction rate of over 50%. This task can be broken down into the two tasks described above: statistical analysis of market data, and statistical analysis of economic agents.

The first problem to tackle is the analysis of market data. This seems like the most natural place to start because it was the first way in which big data analytics were applied to financial analysis. First and foremost, one needs to acquire the data. For financial data of both stocks and crypto markets, many public facing API’s are available for free use. These have stores of historical financial data as well as streams of real time data to utilize. Some examples of these services are Polygon.io, Alpha Vantage, IEX Trading, Tiingo, and Intrino for stock data, Coin API, Crypto Compare, Coingy, and Nomics API for crypto data. These services are plentiful with well-maintained documentation from a simple google search. While it may seem odd that these services are free it makes sense when one thinks back to the Efficient Market Hypothesis. A well-informed market is an efficient market, and an efficient market is a profitable one, thus maximizing the availability of information improves every investors profit.

While there are ways of manually performing analysis on large swaths of economic data, this presents problems whose solutions are more suitable for a financial student to find. Instead, I decided to turn the financial problems into a computer science problem and utilize machine learning to do the heavy lifting. This is a trend that many large financial institutions to small time investors have picked up on and experimented with to great success.

“the networks rapidly adapts to the basic shape of the time series and continues to learn finer patterns of the data. This also corresponds to the Adam learning scheme that lowers the learning rate during model training in order not to overshoot the optimization minimum. After 10 epochs, we have a pretty close fit to the test data!”

This excerpt is from a Medium article in which the author, after attending a hackathon, decided to try utilizing machine learning and market data from the S&P 500 to create a market model. He emphasizes the ease of use of such technology especially for people lacking both an economic and neural network background. The primary task of the programmer when utilizing a machine learning engine is to dictate the number of inputs, number of outputs, size of the neural network and its layers, as well as provide test data and tweak threshold values that effect its computation. Data formatting for the inputs here requires minimal standardization because of the APIs we are utilizing for the data acquisition. The engine then builds a trained market model utilizing historical data. The trained model can then be tested against another data set or real time data to check its accuracy. With the help of python analysis libraries such as Pandas and NumPy can help present and further analyze the performance of the trained model. Adjustments to the input and output thresholds, as well as the size of the neural network can be adjusted based on the analyzed performance to improve the model’s accuracy.

As stated before though, just analyzing raw economic data about a given market is neglecting a large quantity of sentiment information that has a very measurable effect on the market. Press and social media, whether factual or opinionated, are taken into consideration by investors within any given market.

The first step towards accounting for sentiment is acquire the relevant data. Social media powerhouses such as Facebook and twitter both have well documented and free APIs which allows easy acquisition of sentiments. News media sites get a little trickier requiring an investment in fine tuning web scrapers to scrape article text from news sites. This process can be relatively involved even with the proper tools requiring fine tuning for each news outlet. Acquiring the data is already a bit of a hassle, but processing the data becomes even more difficult. While the outputs are simple, referring to an asset in a positive context or a negative context, the input of language with its grammar and syntax can be difficult to navigate.

First and foremost, the data must be standardized to a single form we can replicate operations on. This process is known as lemmatization. After that we will remove all stop words. These are grammatical words that do not contribute in indicating the context of the posts sentiment. Next, we remove all erroneous text such as URLs, links, emails, etc. Finally, we remove duplicated of information. While duplicates can be accounted for to adjust the weight or reach of a particular sentiment, it is a waste of time to process the same sentiment more than once.

Once the data is standardized and processed, the analysis can begin. When processing text posts, utilizing map reduce libraries such as mrjob can do all the heavy lifting. Creating a map of each individual input and post in combination with a carefully crafted dictionary can determine, what asset is being talked about and if the sentiment is positive or negative. This process could also be relegated to a machine learning engine. Pandas and NumPy can be utilized compile and analyze all collected information. One could also segment the sentiment data by trading period and overlay it with economic analysis to see certain sentiments relative impact on stock prices and tweak the dictionary accordingly.

While researching into the domain of big data applied to financial analytics, I’ve gotten a relatively good idea on the type of project I want to tackle. First and foremost is to build a market model with machine learning, while utilizing NumPy and pandas to tweak it. If possible, I would also like to dive into textual analysis of social media. Utilizing twitters API gives me a simplistic data stream that requires minimum data formatting preparation allowing me to focus my efforts on building an effective dictionary or possibly machine learning. This secondary goal is one that I hope to complete but will remain ongoing after the conclusion of the course.