Big Data Analytics Capstone

Methodology and Model Design

Damian Etchevest

St. Thomas University

CIS-627-170

Table of Contents

Introduction

The aim of this study is to determine the relationship between news events and the stock market. To accomplish this, we study the patterns of behavior on days which have a news event and test them against those days which do not have news releases on market hours. We are looking to determine the wave (pattern) on market parameters.

Data Sources

The data sources utilized for stock market data differ from those of news data. For the first ones, we use the API from Dukascopy, which allows us to pull historical data on a considerably number of instruments at a rate of 5000 quotes per call. The API requires an application, but it is essentially free and fast. On the other hand, for the later resource, we utilize various financial new sources including but not limited to Bloomberg, Wall street Journal, and the Financial Times.

As stated before, the purpose of this project is to test if market waves behave significantly different in days of events as on those days of non-news releases. Therefore, top of book market data, which provides order matched data, should be the most beneficial form of source. Nasdaq, NYSE, and IEX represent the list of stock exchanges where our instruments of interest trade. Our research assumes that institutional investors perform mostly on Nasdaq. However, the capital investment that it would require to get top of book market data makes it difficult for readers to replicate the research. Therefore, we utilize tick-by-tick data, which is free and consists on bid-ask, a two-way price quotation that indicates the best price at which a security can be sold and bought at a given point in time, and their respective volumes.

Our research focuses on the behavior of FAANG (Facebook, Amazon, Apple, Netflix, Google) stocks market. The reason behind this decision is because of the high frequency of news release and the volatility on their respective markets.

In order to develop a model that does not overfit, the rule suggests training with a good volume of samples. For this project, we utilize data from market hours in 2019. The dataset was built with a loop, which makes calls on market hours, through every day of that year. The call parameters include key, instrument, timeframe (tick), count (5000 as max), start and end (from 9:30 to 15:56).

Data processing

As stated before, the data pull is an automated batch process which makes calls every five minutes (in order to allow the market accumulate ticks), at a rate of 5000 quotes per call, which is the maximum allowed by the source. At every call, we process the data by developing volume and tick bars. This decision goes in hand with the assumption that time bars produce less accuracy [1], as they intend to produce information at a fix interval (time). Therefore, volume and tick bars provide

both the benefit of synchronize sampling (at a pre-determined threshold of an estimated amount of bars per market session) with a proxy of information arrival, and of normalizing the returns, which is assumed on most machine learning algorithms.

This research assumes that partial information should be carried form bar to bar, by doing so, we focus on market trends while the intention is to avoid bar spikes (outliers). To do so, the last quartile of the ticks contained in the current bar are stored and added to the next bar. Further analysis is required to verify the significance difference of the method.

Several variables are calculated when developing the bars:

- Ticks inside the Bar
- Volume from bid-ask
- Price (different formulas)
- Spread of bid-ask quotes
- Tick Rule (price change from ticks)

Procedure

In order to determine if there is significantly different pattern of behavior developed on market days where a news catalyst occurs, we have centered the analysis on the aggregated tick and volume bars from the different instruments processed before. By doing so, we are certain that any analysis does not overfit a specific instrument, and therefore, be confident to make claims about the results which represent the real behavior of the market psychology relevant to any instrument.

Some relevant assumptions are to be tested in the research, which include that before news release, market movements significantly different from non-news release days express the intentions of informed investors, which might refer to institutional investors. Previous research use methods to find the probability of the presence of informed traders. To do it, they use top-of-book market data to predict on three proxies: the aggressor side of trading(as given by buy-sell indicator flags in the data), an estimate of spreads, and the permanent price effect of trades. Further analysis will apply those methods and test for the accuracy of our model against those. On the other hand, after news release market movements significantly different from non-news release days express the intentions of the public, or retail investors. To address the difference on before/after news release sessions, we develop an index which is 0 at the moment of the news release, or the closest on distance bar.

Another key method implemented involves the development of a function which calculates the normalized (to deal with outliers) absolute (to deal with negative values) percentage change on any specified parameter. These procedure allows us to evaluate index detailed behavior on catalyst days, therefore being able to identify the exact index location of the average wave start, end, and strength.

Future Procedures