Update on Data Sources

Kevin Lu and Jevin Maltais

December 21, 2017

1. Introduction

Over the past several weeks, we have been building out the data infrastructure necessary to build an equity screening and selection tool. The tool is trained on manually labeled data and utilizes machine learning techniques to systematically identify trading opportunities. This report provides an update on the current state of the data infrastructure and the data sources as well as some follow up questions.

2. Note on Free Versus Subscription-Based Data Sources

During the exploratory phase of this project, we are currently using free data sources using a combination of widely available APIs and web scraping techniques. In our previous experience, the data sources we have chosen have high levels of reliability and quality and they are suitable for use in this phase of the project.

Data sources that require a subscription fee, however, offer higher levels of reliability and quality, so as this tool moves to production, you may want to consider subscribing to those data vendors for a small fee. Our recommendation, based on prior experience with working with subscription-based data vendors and evaluating both the reliability and quality of the free data sources we have selected, is to continue using these free data sources while monitoring for any potential problems. The tool can switch to using a subscription-based source later on if necessary.

3. Publicly-Traded Company List

We have compiled a list consisting of approximately 9,000 publicly-traded companies. When the tool is production-ready, it is likely that the tool will be able to identify trading opportunities across the entire universe of these 9,000 companies. This list was compiled by sourcing the publicly-traded companies that are listed on the NYSE, NASDAQ, AMEX, and TSX located on the NASDAQ website and the TSX website. The tool regularly downloads the latest list of companies so that new companies will be included.

The data for US-based equities contain the symbol in addition to market capitalization information, sector, and industry information. The data for Canada-based equities contain only the symbol.

Questions:

1. Are there companies on other exchanges that you would like the tool to consider?

2. Is there any lower bound to the market capitalization or liquidity that needs to be met in order for you to be interested in trading the stock?

4. Pricing Data

The primary source for pricing data for each company is obtained from Yahoo! Finance. The data contains open, high, low, close, and volume information. The secondary source for pricing data is obtained from AlphaVantage which provides similar data in a similar format. For exploratory and testing purposes, we have downloaded data for all US-based equities with a market capitalization of over $1 billion which consists of approximately 2,500 companies.

5. Earnings Announcement Data

Earnings announcement data is obtained from NASDAQ's earnings calendar which in turn sources data from Zack's Investment Research (a subscription-based data vendor). This earnings calendar provides both confirmed and expected earnings announcements. Based on our prior experience and research, we believe that Zack's represents the gold standard for earnings announcement data.

However, even when relying on a subscription-based data vendor, getting perfectly accurate earnings announcement data is hard since ultimately compiling earnings announcement dates requires a manual process. We mention this to provide transparency that the tool may not always handle earnings announcement dates with perfect accuracy since the data that is available to the tool is itself noisy.

If the current data source we are using is proven to be insufficient from a data quality perspective, you can consider subscribing directly to Zack's (which should provide better quality data) or an alternative data provider.

Questions:

1. What is your current process for checking earnings announcement dates?

2. What is the reliability of the current data you are using?

3. We understand that typically you wish to avoid holding through earnings announcements unless the trade has already performed well in the past. In cases where you exit the trade before earnings announcements, how far in advance do you exit the trade? Do you hold all the way to market close when an earnings announcement is scheduled to be released after hours?

6. Next Steps

The next steps we are working on will be to incorporate all the data sources identified along with the manually labeled data you provided us to form a clean dataset to work with. We will also perform some exploratory data analysis and generate some visualisations to verify the integrity of the data and explore the relationships between the data and position labels. Please let us know if you have any additional questions.