Pranjit Kumar Kalita

Budget Proposal for Data Gathering and Subscription

Princeton University, Spring 2017

Introduction

In my intended research, I am interested in understanding the behavior of stocks through machine learning algorithms in order to predict market inefficiencies, alongside using Algorithmic Game Theory to measure individual investor strategies to attain an "equilibrium of strategies" in the future. I believe that if successful, these two components would enable us to place our best trading strategies forward and might further enable us to stay ahead of the markets by noting other market players' actions relative to different scenarios. Therefore, my pursuit will be to reach that state in which we have successfully backtested learning models, and on a macro level, model the various learning models automatically to make our strategy more efficient.

In this document, I will present my findings on various big data public stock sources and present how to best use the given budget of $100,000 suggested by Professor Han Liu. While it's noteworthy that the proposed subscriptions and data gathering will leave a lot of capital at hand, we can use some of the remaining money on gathering high level data of hedge fund strategies and their performance over time, their risk to return ratios relative to particular investments, etc as reported by them, that I suspect will be immensely beneficial in making informed decisions for our fledgeling probable fund as we begin work on it.

I will structure this document according to the following -

1. **Big data gathering sources for publicly traded stocks**: Compare various vendors/outlets and their core competencies relative to each other.

2. **Event-driven structured data gathering**: Compare news-fueled public stock data to examine and extract future market inefficiencies from.

3. **Ease of API usage**: We need to be able to pull and fit structured data into our models with low latency.

4. **Hedge Fund Industry Reports**: This is a source that provides us with an interesting private sneak peek into hedge fund performances as reported by the funds themselves, in areas such as fund strategies, portfolio profiles and performance, investor profiles inside each individual fund, risk to reward ratio of known investments.

5. **Portfolio Management Tools**: This is an important tool that we can use to diversify our intended investments, and is essential from a research perspective since we want to know not only what stocks but how a weighted pool of them would have behaved over time upon backtesting.

Sections 1, 2, 4 and 5 are essential components of our intended research, and they each help either directly or by providing certain frameworks in which we can test our theories on, by allowing us the flexibility of not wasting time designing systems for doing

a lot of our testing operations. They further enable us to try out “aggregate scenarios”, which will be essentially useful in the starting days.

Section 4 is of consequence not directly to the intended research project, but will give us insight into current industry knowns and unknowns, and will only better inform us to pick securities even when picking which models to test them on. Furthermore, we will gain an insight into risk and make us better informed should we be able to commercialize our research.

Big Data Gathering Sources for publicly traded stocks

The following are possible outlets for market data sources of publicly traded stocks -

- **Bloomberg** - Bloomberg has perhaps the best service covering different facets of running a hedge fund - from data to risk and stress testing to integrated trading platforms, irrespective of where the fund is at in its life cycle. Their Bloomberg API initiative (BLPAPI) is an open source tool that allows subscriptions for real time data, requests for historical/reference data, two functionalities of core value to our data gathering. It allows global market data distribution to different kinds of enterprise and desktop applications. As a developer, the market data can then be fed to create proprietary applications. Data is provided through Bloomberg's Market Data Feed (aka B-PIPE) which is collected from over 330 real-time exchanges. This market data can be delivered in conjunction with third-party or Bloomberg distribution platforms for algorithmic trading application. **Cost of subscription to Bloomberg Professional Services - $24,000 / year**. However, Bloomberg has a platform for education called Bloomberg for Education that partners with universities to give all of its market data (both historical and real time) for free. We should explore this avenue. We will then have access to the same data as financial professionals with Bloomberg Terminal subscriptions do.

- **BATS** - BATS U.S. Equities Exchanges provide trade data for market participants looking for real-time market data. However, based on what I've been able to research so far into them, not ideal for historical data. **Subscription cost - $10,000 - $15,000 / month**.

- **Quandl** - Delivers market data from several sources via API, or into Python, R, Excel. Attractiveness of Quandl is that we can receive it in the desired format based on the targeted asset class and can be filtered according to data type and region. Historical data on equities date back to 1996. The API for Python is especially easy to understand. Best part is that the **Quandl Python module is free.**

- **Intrinio** - Provides financial statements, stock prices, news, valuation ratios. Formats - Excel, Google Sheets, Web API (very easy to use). **Cost - Free basic, developers free, individuals $40/month**. However, this company is targeted at developers creating financial apps using market data, but nevertheless seeing their API, it works for our purposes.

- **Thomson Reuters & S&P 500** - Standard. Access to historical and live data of all major markets. Split across several assets classes. APIs provide access to both live and historical data. **Subscription cost for Thomson Financial - $1800/month.**

- **Xignite** - Very similar to Intrinio, focuses on financial data gathering through an easy-to-use API for financial app developers. Heavily focused towards the mobile app industry. **Subscription costs - $3,000 to $13,000 /year**. Also includes a trial run with an API cap of 250 hits in 7 days. Pricing might change relative to the type of data not

needed, based on unrequired asset classes such as bonds, credit, derivatives, etc, not required for our project.

- **FactSet** - Real-time stream data to in-depth historical information. Also integrated with several leading statistical software packages. API provides access to streaming and historical data in CSV format. **Enterprise subscription cost - ~$20,000 - $25,000 /yr**. This is highly used and well-received, so definitely one source we can try to use.

Event-Driven Structured Data Gathering

Using machine-readable news data to find alphas and inform trading strategies. The following two sources were found and will be highlighted in this section -

- **Bloomberg Event-Driven Trading Feed** - subscription cost - $20,000/month.

A. Live streaming textual news data - Not significant for us.
B. Historical Archives - Very important. This segment contains years of historical news related data to backtest models with. Old data in machine readable format can be used to check the viability of alphas and predictive signals to integrate with their trades.
C. News Analytics - The most important segment of the service. This contains machine-readable structured data containing news items and shows their impact in real time. It is subdivided into the following :-
    1. *Sentiment Analysis* - An AI solution showing how each ticker is supposed to have been perceived in the marketplace.
    2. *Readership* - What is the most significant news story at any given point in time. Shows the relative interest and can be tied to create more weightage to our portfolio for corresponding securities.
    3. *Publication Flow* - How a particular security is measured in terms of velocity and supply. Used to highlight events relative to securities.
    4. *Volatility & Impact* - How a story will impact asset prices. Bloomberg leverages proprietary data sets to perform large-scale regression analysis and predicts impacts. A Machine Learning solution.
    5. *Novelty* - measure of the likelihood that a story is introducing new information to the market. Traders can adjust their signal weights accordingly.

Seeing that we will have a surplus relative to the intended budget, we can certainly avail this subscription for one month to check the reliability and strength of our learning algorithms within this exciting “new-world” trading framework. I strongly believe that the ability to create arbitrages from sentiment analysis is the future if not present already, and is only going to get more lucrative with time.

- **RavenPack Asset Management** - Provides systematic analysis of unstructured big data sets, such as traditional news and social media by transforming them into granular data and indicators. In their website, they say that funds use this service for the following - generate more alpha, risk management, generate trading ideas and cutting false positives in market surveillance.16 years of historical data for backtesting is provided along with 30 days of free trial for financial institutions. **Subscription cost is $10,000 /month**. I believe this is a very reliable service too, and can be reached out on behalf of Princeton University, to which they will probably agree to give us subscription.

In terms of how it works, RavenPack indicators provide sentiment analysis for an entity that could be used alongside other technical indicators such as risk/reward history

to build updated portfolios. Refer to the paper - “A Machine Learning-Based Trading Strategy Using Sentiment Analysis Data” by Lucena Research, Feb. 25, 2015.

Ease of API Usage

The good news is that all of the data sources highlighted in the first section are supported by very competent, low latency APIs for real-time data streaming, as well as are equally compatible for querying of historical data. As noted before, Quandl's Python package which is a free and very intuitive module is the most cost-effective at this point of our life cycle. The following sections will give a brief overview of each of the other data outlets and their corresponding APIs :-

- **Bloomberg** - Multiple programming languages like C, C++, Java, C# supported. Securities identified by id, and with a subscription id, can pull requests fairly easily. Historical and intra-day time series data. The API is a client-server framework that allows non-Bloomberg terminal users to pull data from Bloomberg servers. Capabilities like caching aid in faster retrieval. Several programming instances from their website makes me confident of their easy-to-use nature.

- **FactSet** - Supports C++, C, Java for use with real-time data feeds. Allows access to historical time and sales data collected from exchanges on a daily or ad hoc basis. Specific time intervals could be specified. Data returned in CSV format, which can then be manipulated for our own applications or learning algorithms. Requests and receptions are considered secure and follow industry-standard web protocols.

- **Xignite** - Supports Java. Allows free-trial period for API. Java SDK looks very approachable and intuitive. Flexible pricing model based on asset class desired, and frequency of data requests.

- **Quandl** - Python module runs on either Python 3.x or 2.7. Easily downloadable from Github. 50 API calls/day is the limit.

- **BATS** uses a version called XigniteBATSRealTime API, which is powered by Xignite mentioned above. Not ideal for historical data collection.

- **Intronio** - Uses a web API. Not programmable, file downloaded in CSV or Excel format. Different access codes for different data feeds. Good for historical data. Can slice data based on industry, company, time period. However, there is a minimum capital a startup should have raised to gain subscription, so not sure if this will be available for educational setting.

Hedge Fund Industry Reports

A company called Preqin has extensive self-reported private information of hedge funds. It is a great resource to understand the following -

- Investor Profiles for each fund
- Hedge Fund performance - Provides a great market overview. Compare funds to specific market segments, sorted by strategy employed, fund structure, geography, etc.
- Investor Portfolio Performance - Risk/Reward of known investments.

I took the liberty to obtain a demo from them, where they showed how their system is structured. Specifically with regards to risk/reward of known investments relative to strategy, they showed an example (.pdf enclosed in separate document) of Bridgewater's All Weather Strategy with active components including -

- Monthly returns over past 10 years
- Annualized performance sliced over different year periods
- Benchmark Strategies dating back 10 years
- Evolution of AUM
- Rolling Volatility levels over the past 10 years
- Sharpe Ratios over the past 5 years

I believe having access to this data is helpful in a high level way to our fledgeling project in the following manners :-

- Become well-versed with the scale of activity undertaken in the industry
- Understand in real world how theories and theoretical concepts work (eg: Sharpe Ratio)
- Gain insight into their strategies relative to investments
- How do industry leaders do well relative to their competition
- Gain important insight about what investors to target

The **subscription cost is $14,100/year** and it includes access to up to 5 different people within the same subscription setting. I hope you will find this tool useful. They also provided me with a free trial period and I will be doing further in-depth homework into their database during Spring Break. I believe we should get access to it.

Portfolio & Risk Management Tools

Portfolio & Risk Management Tool, provided by Bloomberg, is a management tool that allows us to analyze a portfolio's historical performance using transaction or factor based attribution, to evaluate portfolio structure and monitor daily performance in real time. We would be able to implement optimal investment portfolio strategies using this tool. Enabling the testing of hypothetical situations with trading simulation and portfolio optimization provides a comprehensive understanding of potential sources of risk, through a multi-factor risk model. This service is part of Bloomberg Professional Service, with subscription cost outlined in the first section (comes with the Bloomberg Terminal at no additional cost).

Some features :-

- Past, present and future positioning of a portfolio
- Position analysis and investment simulation tools to analyze portfolios
- Performance evaluation with custom stress tests
- Assess portfolio risk factors
- Customizable relative to set benchmarks of each portfolio - historical Sharpe ratios, standard deviation, etc.
- Sector-by-sector comparison of portfolio performance relative to benchmark

As noted, this tool is part of Bloomberg Terminal, and if we were to get it, we could perform stress tests to measure risks relative to different portfolios. At first sight, it wouldn't seem necessary since we are a machine learning and predictive analytics shop, but if we could integrate predictions with stress tests relative to how several investments performed in the past, we could adjust our investments accordingly. This tool would be a good tool to have, and I'd be happy to discuss more about its use in our project.