**CIS 545 Course Project Proposal**

By Steve Brooks, Chenlia Xu

November 2021

1. **Team members**

Steve Brooks

Penn ID:

Email: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Major responsibilities:

* Collecting historical stock trading data
* Run linear regression model and interpret results
* Construct portfolio and trading strategy

Chenlia Xu

Penn ID:

Email: [chenlia@seas.upenn.edu](mailto:chenlia@seas.upenn.edu)

Major responsibilities:

* Collecting all other data, join them with stock data
* Data wrangling and cleaning
* Run PCA to decide the components
* Add / remove features

1. **Data Sources**

We planned to use a combination of the following data sources:

* JoinQuant for historical stock trading data
* US statistic census for economic data
* The World Bank public database
* … anything else?

1. **Project Plan**

The intention of this project is to use various sources of data to analyze their impact to individual stock performance so that we can develop a trading strategy to construct and maintain a portfolio that can outperform the market indices.

We expect the data to come from various sources, including:

* historical stock trading information, e.g. price, volume;
* company financial performance, e.g. revenue growth, earnings;
* analyst coverage and recommendations;
* economic data, e.g. GDP growth, employment rate, retail sales, real estate prices;
* non-economic data, e.g. power usage, population, CO2 emission, democrats / republican in white house;
* headline news of major economic media, e.g. Bloomberg, CNBC;
* other seemingly unrelated information, e.g. weather, weekday.

We planned to use the models we learnt from this course. It will be supervised learning with historical share price movement as dependent variable. We care about daily / monthly / yearly share price % up/down for individual stocks, as the strategy will be based on projection of future movement to adjust portfolio weight accordingly.

We will likely use linear regression model as we are projecting the stock movement range, not just a categorical direction of up or down. We will apply PCA to pick the most powerful features. When training the model, we will divide the data into training set and testing set for cross-validation.

The ultimate objective is to develop a trading strategy that tells us which stock to buy or sell and by how much quantity. The strategy is supposed to produce a portfolio that will deliver an absolute return higher than market indices, e.g. S&P, over a reasonable period of time. We will also consider other risk factors, like SHARP ratio, maximum retreat.

1. **Why is it interesting**

The interesting thing here is if we can really develop a model that consistently produces great outcome, we can make millions of dollars by applying it in real world! There’re already many hedge funds, quantum investment funds out there applying mathematic model and deep learning in real practice, and some of them made quite decent return. For example, \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_. While we do not expect our model to be perfect or near those pioneers, we want to apply what we learnt in this course to feel about what we can do with the data to generate sensible outcome.

1. **Potential Challenges and Obstacles**

Stock performance is perhaps “unpredictable”. There’re theories around random walk, efficient market, etc. that says one will never be able to beat the market with market available public information. However, these are academic studies and let’s see what data science can do to prove them wrong.

Another challenge we can foresee is that there’re just too many variables that can affect stock performance. Practically we may not be able to gather all the information as our input. Even we can, there’re just too many features and our PCA will result in a very flat cumulative explanation curve.

Also, we will need to introduce transaction cost. In real world, it could be costly to trade on daily and even hourly basis. The transaction cost will drag down our return meaningfully. Liquidity will also be a constraint whether we can easily buy or sell large block of shares. Short selling is not always available for any stock.

1. **TA Support**