Northeastern University, Khoury College of Computer Science

DS5230 - Project Proposal

Dharun Suryaa Nagarajan - MSDS - Semester II
Rohan Benjamin Varghese - MSDS - Semester II

Description of the Problem:

- The 'Pairs-Trading' strategy is popular in hedge funds for minimizing risk and maximizing returns, thanks to its simplicity and market-neutral qualities.
- Pairs trading relies on monitoring correlated stock pairs, opening long and short positions based on historical behavior to profit from convergence in the long run. Unsupervised learning algorithms are used to identify these pairs.

Summary of the data:

There are 3.7M data points available across 96 variables from CRPS data sourced from the WRDS website. Some key variables are the following:
absacc: difference between a company's reported net income and its cash flow from operations.
invest: the sum of a company's capital expenditures and inventory investment.
cashdebt: This metric represents a company's ability to service its debt obligations.
dy: compares a company's annual dividend payments to its market capitalization.
ep: a valuation metric that compares a company's earnings per share to its market capitalization.
gma: a company's profitability before accounting for certain expenses, such as interest and taxes.
Return on assets (roa): a company's ability to generate profits from its assets.
Return on equity (roe): a company's ability to generate profits from its shareholders' equity.
Sales growth (sgr): change in a company's sales revenue over a period.
MOM1M: Price difference between start of the month and end of month vs at the start of the month.

This is not an exhaustive list of all the metrics included in the table. The table contains many other metrics that can be used to analyze a company's financial performance and condition. It is important to note that no single metric is a perfect measure of a company's financial health. Investors should consider a variety of metrics when evaluating a company.

Methods:

- Since our dataset is a very high feature dataset, we are planning to reduce the number of features using PCA (linearly) or in a non-linear fashion using CAE(convolutional AutoEncoder technique).
- Next step would be to cluster the underlying risk factors using the best performing algorithm out of K-means, Agglomerative and DBSCAN based on evaluation metrics. We would also like to use reinforcement learning based pairs trading as suggested in the paper to generate training signals else we will be using mean reversion if time permits.

Preliminary results:

- 29829 unique stock indexes are present in the dataset.
- The date range is from 1957 to 2016 with distribution relative to the age of the stocks. From the frequency distribution DATE wise, the number of stocks before 1980 is negligible.
- Around 9k stock indexes have less than 48 months of data which might be less helpful for the pairs trading strategy. 75 columns have more than 15% NULL Values in them. They should be imputed or dropped from the dataset density of the NULLs.
- Although we don't have a method to convert PERMNO to the actual stock and calculate its price, MOM is a good enough proxy, and provides the change in price for that month.
- We are using 1 to 64 window sizes, it will allow the models to capture both short-term and long-term trends. While clustering is distance and hierarchy based, we do want to add some temporal dynamics allowing the clustering algorithms to consider the time series behaviors.

References:

1. Roychoudhury, Raktim and Bhagtani, Rahul and Daftari, Aditya, Pairs Trading Using Clustering and Deep Reinforcement Learning (July 9, 2023).
2. Shihao Gu, Bryan Kelly, Dacheng Xiu, Empirical Asset Pricing via Machine Learning (May 5, 2020).
3. Chulwoo Han, Zhaodong He, Alenson Jun Wei Toh, Pairs trading via unsupervised learning (September 28, 2022)