Thomas Parmentier
Thomas.parmentierlux@gmail.com

CAS in Advanced data science – Module 5 – Peer consulting report of data science project: How to select stocks within the S&P 500 for pair trading

Objectives of the project:

This project aims to select specific stocks within the S&P 500 for pair trading. Pair trading is a trading strategy that involves matching a long position (buy a stock) with a short position (sell a stock use as funding for the long position) in two stocks with a high correlation. This project uses machine learning to select stocks with same characteristics that could be eligible for a pair trading strategy. The stocks are then tested for cointegration to check if a relationship exists between those stocks. Finally the project use portfolio management libraries to assess if the strategy is profitable.

Dataset & methodology used to perform the analysis

The dataset uses a 4-year time period using daily data from the S&P500 through the Yahoo Finance API. The dataset has more than 505 stocks for 1008 days. The dataset has more than 500K observations.

Steps realized before applying ML methods:

- The project is checking for NA and missing value and remove any missing observation After cleaning the dataset contains still 502 stocks and 732 days which is a long enough to pursue the analysis.
- As a second step the daily price change is calculated for each stock and annualized. The same is done for the volatility using the standard deviation.
- As stocks can generate very different returns and volatility, we need to put them on the same scale by using StandardScaler().

Once the three steps realized ML methods can be applied to the dataset.

- The elbow and silhouette methods are used to find the number of clusters
- K-means clustering is applied to the dataset once the number of clusters is selected
- A second methodology is used which is Hierarchical clustering
- Apply the cointegration analysis to the different clusters to find out pair trades
- Use the TSNE to visualize the data in 2D
- Calculation of the performance and back testing of the strategy

What you did right?

- The dataset is well cleaned and analyzed using various libraries such as «missingno», which allows for an easy visualization of missing data in a large dataset. This library is particularly well fit for this project which has a wide range of observations.
- You are using different methods to define the clusters, which gives more confident that clusters are not «forced». The sure of two clustering methods to cross-checks if the results are different is also a good idea and gives more confidence in the robustness of the results of this analysis.
- The project uses lot of visualization, which is helpful to understand the dataset and the different steps performed. The use of the TSNE to visualize the results of the analysis gives an easy and understandable way to understand the output of the project for somebody with no knowledge in financial markets.
- You are using libraries and methodologies used during the different modules of the CAS. For example, some statistical methods seen during module 2 (cointegration) and also ML methods seen during module 3 (K-means, Hierarchical clustering)

What could you improve?

- Extend the dataset to other indexes, could we see meaningful results with the Dow Jones? CAC 40? Or is the strategy only working for the S&P 500.
- The K-means/Hierarchical clusters can be difficult to differentiate in the model. Use more deep learning methods to check if the result would be different.
- On the top of stock prices, you could add more fundamental data, for example valuations, balance sheet or income statements. It could give more differentiation between the clusters and the elbow curve could give a clearer result on the number of clusters.
- Try other normalizations/standardization methods to see if the results look the same
- Use a broader time frame if possible as the relationship between stocks can change over time.
- Put more comments in the code, in order to better understand the workflow.