Citi Code Challenge 2022

Approach Paper

College Name: **MIT Academy of Engineering.**

Team Name: **Optimistic Minds.**

Overview

➢ Describe your understanding on problem statement

Pair trading is basically a strategy in which we can bet on pairs of stocks which will diverge or converge in price. The price of the two stocks needs to be similar or at least close to one another. But at times when there is divergence or convergence in these prices, the trader can decide to go long or short on one of the stocks. There are many other indicators or factors that affect the choice of this pair of stocks.

The problem to be solved here is to generate a stack rank which basically ranks all the pairs based on the profit that can be obtained from them. The highest rank will have to be the pair that is highly correlated and thus can yield maximum profit.

We have to create an efficient model with an UI that can help us generate this stack rank. The model has to be created using the provided dataset. The dataset of the stocks contains the name of the company, date and close price of the stock which helps to identify which strategy should be used to obtain maximum profit in pair stocks. Indicators identify the performance of the stock in the market which helps to gain as maximum profit as possible. But there are numerous indicators which are used to analyse the performance of stock in the market, so which indicator should be used to gain maximum profit is decided according to the current market conditions. So we have to implement a model which will give a pair of instruments that gives leading correlation to maximise profit.

➢ Team members

1. Onkar Chendage
2. Nandini Barkul
3. Shwetali Desai
4. Pratham Madhani

High Level Solution Approach

➢ Describe Solution

Approaching the solution:

1. Preliminary Data Processing.

After considering the completeness and variety of data. The pairs trading requires a mutual economical correlation between two instruments, data pre-processing is necessary, so that the trading performance can be easily evaluated with less bias.

2. Correlation:

Correlation is quantified by the correlation coefficient ρ, which ranges from -1 to +1. The correlation coefficient indicates the degree of correlation between the two instruments. The value of +1 means there exists a perfect positive correlation between the two variables, -1 means there is a perfect negative correlation and 0 means there is no correlation.

A perfect positive correlation is when one variable moves in either an upward or downward direction and the other variable also moves in the same direction with the same magnitude.

Whereas a perfect negative correlation is when one variable moves in the upward direction and the other variable moves in the downward (i.e. opposite) direction with the same magnitude.

According to pair trading strategy we need perfect negative correlation.These pairs having negative correlation would further be used for cointegration.

3.ADF TEST:

The key of finding valid pairs is to find the cointegration of two selecting stocks. As we will go in detail later, we want to find two stocks that their time series of prices follows a linear relationship but not always. To find such pairs, we will perform ADF test (or Augmented Dicky Fuller Test) to every pair in each cluster to find cointegrated pairs. ADF test is usually used in time series analysis. In this case, the ADF test helps us determine whether the spread of two stocks is stationary or not. A stationary process is very valuable to model Pairs Trading strategies. The ADF test gives p-value as the result.

4.Cointegration:

The parameter cointegration is used for determining the statistical connection between two instruments. Two-step method can be utilised and the following hypotheses will be tested between the instruments, which represent the two stock prices.

H0 : There is no cointegrating relationship;

H1 : There is cointegrating relationship,

To show that two stocks are cointegrated, the null hypothesis should be rejected.

Study has set 0.05 as a threshold to screen out all the cointegrated stock pairs, and the asymptotic p-values were calculated based on the MacKinnon's approximation used in the Augmented Dickey-Fuller unit root test.

When the p-value is smaller than 0.05, it indicates strong statistical evidence against the null hypothesis, and the alternative hypothesis should be accepted.

5.Z-score:

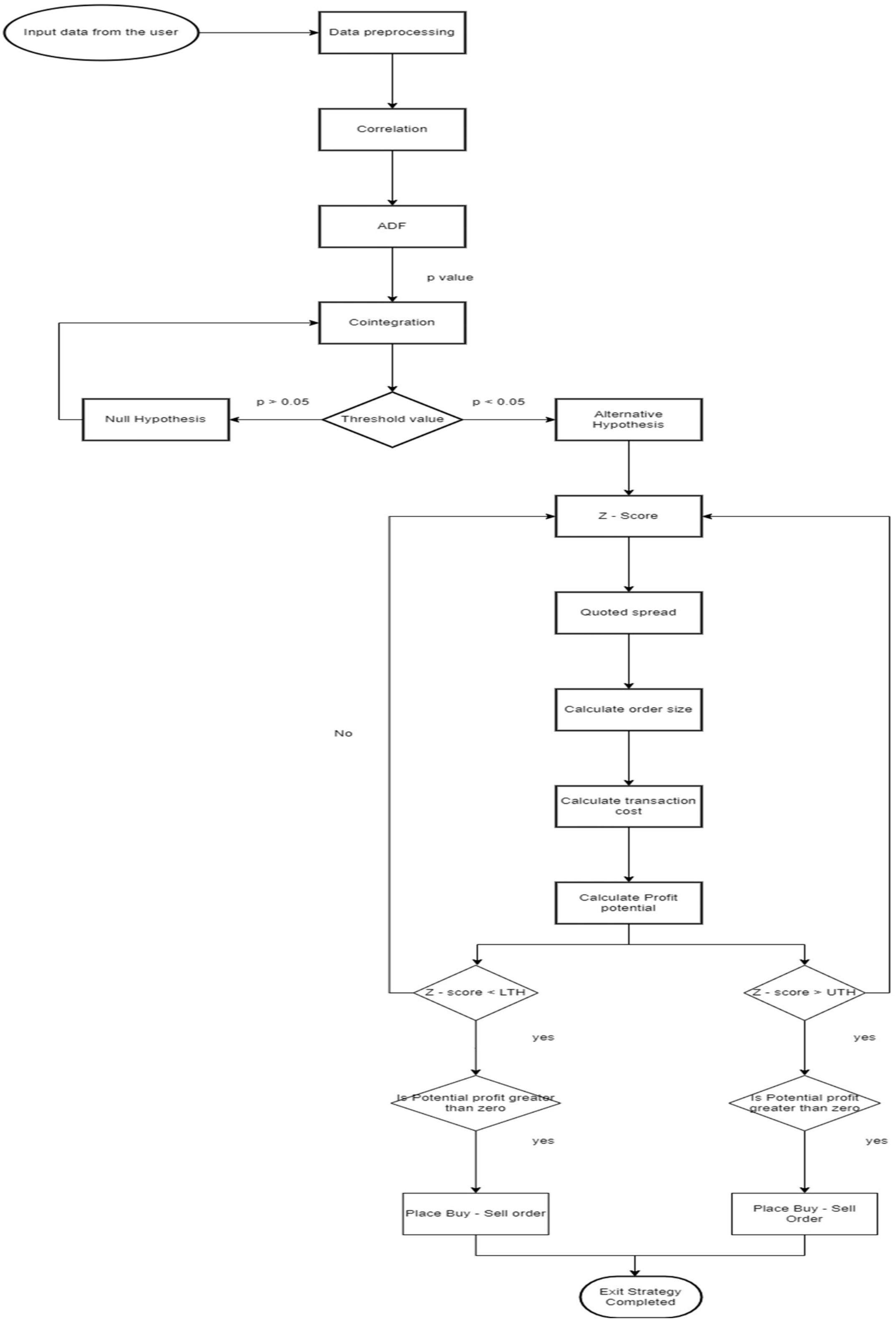
Z-score is used for trading strategy. Z score value is high or low decides stocks are sold or bought. When making a trade, the actual ratio does not give precise statistical information.

Instead, the relative movement of the ratio should be studied. Given a normal distribution of raw data points, the z-score is calculated so that the new distribution is a normal distribution with a mean of 0 and a standard deviation of 1. Having such a distribution ~ N(0, 1) is very useful for creating threshold levels.

6.Backtesting:

We apply our trading strategy to the real stock market and check how much we can earn based on our approach. We used the moving windows approach for the testing. The input of back testing is the z-score history generated in the 'trading strategy' part and the price history. Based on the input, we keep calculating the earning and loss of our stock and inverse. We also track the total asset history and return it as an output of back testing.

➢ High Level Architecture



Flowchart Image Link:

https://drive.google.com/file/d/1BYaSQJSmW24YF6dfdLrHx7e1YzUKEXA0/view?usp=sharing

Impact

Product Approach

To illustrate the potential profit of the pairs trade strategy, consider Stock A and Stock B, which have a high correlation of 0.95. The two stocks deviate from their historical trending correlation in the short-term, with a correlation of 0.50. The arbitrage trader steps in to take a dollar matched the long position on underperforming Stock A and a short position on outperforming Stock B. The stocks converge and return to their 0.95 correlation over time. The trader profits from a long position and closed short position. For example: Pepsi (PEP) and Coca-Cola (KO) are different companies that create a similar product, soda pop. Historically, the two companies have shared similar dips and highs, depending on the soda pop market. If the price of Coca-Cola were to go up a significant amount while Pepsi stayed the same, a pairs trader would buy Pepsi stock and sell Coca-Cola stock, assuming that the two companies would later return to their historical balance point. If the price of Pepsi rose to close that gap in price, the trader would make money on the Pepsi stock, while if the price of Coca-Cola fell, they would make money on having shorted the Coca-Cola stock.

Business Impact

Pairs trading is profitable. The broad market is full of ups and downs that force out weak players and confound even the smartest prognosticators. Fortunately, using market-neutral strategies like the pairs trade, investors and traders can find profits in all market conditions. The beauty of the pair's trade is its simplicity. The long/short relationship of two correlated securities acts as a ballast for a portfolio caught in the choppy waters of the overall market. Business would always on the profit side as there are two trades involved, even if one stock performs in an unexpected way the other stock can make up some of the losses.

Non-functional Requirements

- ### Scalability

  Scalability is the ability to handle an increase in the workload without impacting the performance, or the ability to quickly expand the architecture. A Black Friday Test should be performed. Black Friday refers to the day after Thanksgiving and is symbolically seen as the start of the critical holiday shopping season. Stores offer big discounts on electronics, toys, and other gifts. Stores offer big discounts on electronics, toys, and other gifts. The solution must allow the hardware and the deployed software services and components to be scaled horizontally as well as vertically. Horizontal scaling involves replicating the same functionality across additional nodes; vertical scaling involves the same functionality across bigger and more powerful nodes.

- ### Throughput

  System should be capable of handling a given number of transactions in a given unit of time. For example, a thousand a day or a million.

- ### Security

  Authentication: System should be able to do correct identification of parties attempting to access systems and protection of systems from unauthorised parties

  Authorization: Mechanism required to authorise users to perform different functions within the systems

  Encryption: All external communications between the data server and clients must be encrypted

  Data confidentiality: All data must be protectively marked, stored, and protected

  Compliance: The process to confirm systems compliance with the organisation's security standards and policies.

- ### Entitlement

  Django (free of charge)

  AWS (cost saving)

  PyCharm (Licensed)

  Jupyter Notebook (free of charge)

Cloud Deployment

Cloud deployment is the process of deploying an application through one or more hosting models—software as a service (SaaS), platform as a service (PaaS) and/or infrastructure as a service (IaaS)—that leverage the cloud. This includes architecting, planning, implementing and operating workloads on cloud.

With an effective cloud deployment model, the organisation achieves numerous benefits such as faster & simplified deployments, cost savings, Integration, Usability, Performance, maintainability, scalability, agility. operational efficiency. We will be using Amazon Web Services (AWS cloud) to achieve the solution.

Test Automation

The response time of the application should be verified i.e. how long does it take to load the application, any input given to the application provides the output in how much time, refreshing the browser, etc.

Throughput should be verified for the number of transactions completed during a load test.

Process activities like import & export of data, any calculations in the application should be tested.

ETL time should be verified i.e. the time taken in extracting, transforming and loading the data from one database to another.

Increasing Load on the application should be verified.