Pair Trading Models

Data Analytics Project

1st Hritik Shanbhag
PES University
Bangalore, India
Hritik.24shanbhag@gmail.com

2nd Varun Seshu
PES University
Bangalore, India
seshuvarun@gmail.com

3rd Manas V Shetty
PES University
Bangalore, India
manasvshetty@gmail.com

4th Sashwath S Kumar
PES University
Bangalore, India
shashwath457@gmail.com

I. Introduction

The idea behind pairs trading is simple: find a pair of stocks which track each other, and when they diverge, buy the lower one and short the high. If they converge again, you pocket the difference. It is perhaps this simplicity that has attracted the attention of so many of Wall Street's quantitative minds. First explored in the 1980s by a group of applied mathematicians and scientists led by Nunzio Tartaglia, automated pairs trading was initially a great success, helping the team generate about $50 million of revenue for Morgan Stanley in 1987. The group disbanded in 1989 after a few years of less than satisfactory returns, but many high profile quantitative firms, such as D.E. Shaw and Long Term Capital, have since tried their hand at the strategy, hoping that, with a minor tweak here or there, they could capitalize on the algorithm's early promise.

Pairs Trading can be called a Mean Reversion Strategy where we bet that the prices will revert to their historical trends.The first step in designing a pairs trade is finding two stocks that are highly correlated. Usually that means that the businesses are in the same industry or sub-sector, but not always. [2]

Though pairs trading sounds intuitive, in applying the strategy there are three main problems:

1) How does one identify profitable pairs, that is, pairs which will continue to track each other out of sample?

2) Assuming you can close your position exactly when prices cross again, it is easy to see that your profit the spread between the stocks when your long-short position. It follows that to maximize profits, one must initialize at peak divergence without seeing into the future how does one predict when this will be?

3) In the event that a pair does diverge, how can one detect it and react accordingly

In this project we explore and attempt to answer these questions using data collected over the past 20 years(240 months).We pulled data from Yahoo Finance, which provided us with open and close prices at each day and other meta data regarding the specific company.We try to come up with a model which, given the stock price data, identifies pairs suitable for trading. We go about this by using Linear Regression,Decision Trees and Bollinger Bands.

II. Other Literature in this field

Pairs trading involves finding two stocks whose price movements have been highly correlated, and then, when the correlation is perceived to have weakened, taking a long position in one and a simultaneous short position in the other. [1]

While pair trading we have to assume market neutrality, pair traders expect that the under performing stock will eventually return to neutral performance that is an increase in price. Meanwhile, the same assumption for the over performing stock indicates that a price decrease should occur.

The approach used while using pair trading strategy is:-

First find two stocks that have historically demonstrated a strong degree of correlation in price movement. Stocks in unrelated industries tend to have weaker correlations than those in the same industry. For example, if RBI increases the interest rates, then all corporate banks would be affected the same way and likewise when the rates are lowered and therefore, are likely to be more correlated than across industry stocks.

Secondly then when the correlated pairs are shortlisted , one calculates the historical spread.One should plots indicators or overlays like Bollinger Bands are also applied.

Once we have the ratio plotted, we look to see if the ratio is mean reverting. If indeed it is, then one shortlists this as a likely pair to trade when a divergence occurs.

Finally we involve finding opportunities to trade! .Essentially after plotting all of the correlation graphs ,bollinger bands it is now required to spot times when the ratio has moved significantly away from the norm, offering a trading opportunity. These may be intra day let's take example of two automobile colonies bajaj motor and hero motor so a sudden blip has caused Hero Motor to fall sharply while Bajaj motor is not moving instead of going long Hero, which may be risky, one may Long Hero and Short Bajaj if one presumes the gap will close during the day or positional where one is betting Hero Moto has more momentum than Bajaj or fundamental triggers.

Limitations are:-

Position sizing – When you are attempting to work on price ratios or spreads, you are indirectly assuming that you have equal exposure to both stocks but that may not be true. Example, Hero motors(long) is priced at Rs 3000 and has 500 shares per lot versus Bajaj motors(short) which is priced Rs 2700 but same number of shares , in that case,your trade is not even on both Long and Short so ideally you should adjust for this Rs 300 difference by using more lots till both become equal or by borrowing Bajaj shares in Securities Lending and Borrowing.

Co-integration : In this the two entities like bajaj motors and hero motors seem to be correlated but it doesn't mean they are as there could be something else which was driving this similar movement and that something else may not happen in the future and the breakdown in correlation may be permanent.

As a summary in the field of pair trading it is a popular trading approach that appeals to many, especially to those who like the idea of having a more market-neutral strategy. Just as with any strategy one suggests that the trader have a trading plan with predefined entry and exit rules and have the discipline to follow them no matter what the market does.

III. Proposed problem statement and how we intend to address it

Problem Statement: Find the best stock pairs in the stock market and implement successful Pair trading strategies on those pairs.

We intend to find the best correlated and cointegrated stock pairs in the Indian Stock Market from the years Oct 2017 to Sept 2020 (this is the period we have picked for checking correlation between stocks). We are accomplishing our task in 4 Phases - Collection of data, Preprocessing, Models, Backtesting.

We have scraped our data from the Yahoo finance site, our data consists of historical Indian stock market candlestick data from years 2000 to 2020 of all companies currently floating shares in the market. The data we have scrapped is mostly clean and accurate, but with a few missing values and lesser amount of data for a few companies(as they may have gone public more recently).

We start processing our data by first handling the missing values. We initially tried to interpolate the values of the missing stock data but found that this was introducing inconsistencies in our data and was introducing volatility to stable stocks. Hence, we decided to simply remove rows with missing values as it consisted of less than one percent of our data and would not affect overall correlation of the stocks.

Next, we have removed the stocks with less than 2 months of data as such low amounts of data might lead to spurious correlations. We also limit the data of other stocks to two years as it would aid in faster calculations as the number of possible pairs of stocks is very large and might lead to excessive compute time for calculating correlations and cointegration testing.

To further decrease compute time, we use fundamental data collected from stocks, such as stocks in the same sectors are affected by the same news and can thus show correlation in movement due to changes that affect the entire sector. [3]

We next add the Bollinger Bands for all the stocks in order to compare the volatility of the different stocks. Calculation of the bands includes calculation of the 20 Day Moving Average for all companies closing prices along with the 1, 2, 3 standard deviations both above and below the 20 day moving averages.

We then go about creation of all possible pairs in the stock market, for each pair we create we calculate correlation. We then take the pairs which are highly positively or highly negatively correlated (The outliers of all the pairs formed in the market). We then run cointegration tests on these pairs to help further rule out spurious correlations. We also include looking at multiple correlations for the same stock pair - such as 20 day moving average, standard deviations, opening, closing, high and low prices for each day.

After the pairs are created and the best paris in the market are found, we calculate the spreads between the best pairs in the market using the equation:

$$A - R.B = W \quad (1)$$

Where W is the spread of the stock pair, A is price of stock A in the pair, B is price of stock B in the pair and R is the ratio between them. [3]

We calculate the value of R using linear regression by setting the value of W (the spread to 0). We do this to 'center' the stock around the zero value and measure the deviations

from its value. It also aids in visualizations. It also helps serve as a 'trigger' to make BUY/SELL signals.

We further use these spread values as training data for the decision tree model which uses the spread between stocks to classify whether to buy/sell or do nothing in a particular day given the spreads between the pairs.

Finally, we build a back tester. It helps to test the efficacy of our models. A backetester calculates profit, loss, sharpe ratio and other useful tools for data analysis given the buy and sell orders of the decision tree and it can be used to improve the performance of the decision tree and other models that use the stock data. It also provides us with metrics to compare and evaluate different models.

IV. Different Approaches to solve this problem

Pairs trading is a representative market-neutral trading strategy which simultaneously longs an undervalued stock and shorts an overvalued stock. This strategy is a form of statistical arbitrage trading that assumes the movements of the prices of the two assets will be similar to previous trends.

It follows the assumption that asset prices will return to the long-term equilibrium. This strategy started from the idea that arbitrage opportunities exist when the price gap between two assets expands to or past a certain level. It is also based on the belief that historical price movements will not change significantly in the future.

There are a number of ways to approach this problem:

Reinforcement Learning The fundamental of reinforcement learning consists of two main components: agent and environment.

The environment is represented by different states with a predefined state space, while the agent learns a policy determining what actions to perform out of the action space.

In a full reinforcement learning problem, the learning cycle of an agent could be summarized into the following phases:

1.Make observations of the environment state.

2.Perform action accordingly based on the existing policy.

3.Receive the corresponding reward attributed to the action performed.

4. Update the policy.

Supervised Learning: In supervised learning, the algorithm learns from instructions.

Every instance has an estimation target to compare in order to calculate the cost of discrepancy, and the algorithm is updated by minimising the cost through iteration, so the process is somewhat "instructed" by the target output which tells what is the correct outcome.

Random Forest:

The main idea behind random forest is to reduce the variance of the decision tree models without increasing bias by using bagging. Random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the prediction accuracy and control over-fitting.

V. Our Approach

We are using a decision tree. It is a support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

It is one way to display an algorithm that only contains conditional control statements. We are using decision trees because they help us evaluate between options. What we are looking for is for an output telling us whether to BUY, SELL or not do anything which is FLAT.

We are not using random forests and have chosen decision trees because decision trees are much easier to interpret and understand. Since a random forest combines multiple decision trees, it becomes more difficult to interpret. Also, Random Forest has a higher training time than a single decision tree.

Since our data set is huge it is best we opt for decision trees to get optimal results within reasonable time. We are not using supervised or RL because they introduce neural networks which is out of the scope of this project.

References

[1] Sandeep Raichura, (2018) "What is pair trading and how does it work", 1st December 2018

[2] Uyumazturk, Bora, and Vasco Portilheiro. (2017). "Rise and Fall: An Autoregressive Approach to Pairs Trading."

[3] Papadakis, George Wysocki, Peter. (2007). "Pairs trading and accounting information"