DSSR

Discussion Paper No. 127

Identify Arbitrage Using Machine Learning on Multi-stock Pair Trading Price Forecasting

Zhijie Zhang

July, 2022

Data Science and Service Research
Discussion Paper

Center for Data Science and Service Research
Graduate School of Economic and Management
Tohoku University
27-1 Kawauchi, Aobaku
Sendai 980-8576, JAPAN

Identify Arbitrage Using Machine Learning on Multi-stock Pair Trading Price Forecasting

Zhijie Zhang

Tohoku University, Graduate School of Economics and Management

Abstract

Aims: Market neutral pair-trading strategy of two highly cointegrated stocks can be extended to a higher dimensional arbitrage algorithm. In this paper, a linear combination of multiple cointegrated stocks is introduced to overcome the limitations of a traditional one-to-one pair trading technique. Methods: First, stocks from diversified industries are pre-partitioned using clustering algorithm to break industrial boundaries. Then, cointegration test is performed within each cluster. Last, a linear combination of those cointegrated stocks will be formed using ElasticNet algorithm boosted by AdaBoost algorithm. Results: All three indicators on price prediction chosen for performance evaluation increased significantly. MSE increased by 32.21% compared to OLS, 37.06% increase on MAE, 37.73% improvement on MAPE. (Portfolio return performance is still under construction, indicators including cumulative return, draw-down and Sharpe-ratio. The comparison will be against Buy-and-Hold strategy, a common benchmark for any portfolio)

Keywords: Pair-trading; Arbitrage; K-means++; AdaBoost; ElasticNet; Cointegration

Introduction

As the improvement with computational power and financial instruments, various portfolios and computation-based quantitative trading algorithms are introduced to the applicable industry. Recently, Pair-trading has been one of the most famous trading methods amongst others due to its unique characteristics. Pair-trading is one of the most classic quantitative trading methods being introduced by Morgan Stanley in the mid-1980s. It is also considered to be a statistical arbitrage strategy[1]. The core idea is being that first, select two stocks that have similar historical performance with the assumption of they will continue to do so. Second, trade on their temporal fluctuations where short on the positive movement and long on the negative, the convergence will eventually lead to the original price ratio where a profit can be expected.

Despite the simple ideology of Pair-trading strategy, the advantages of such are outstanding. It is a market neutral strategy with low risk and high stability. However, the core difficulty of finding stocks with similar movement is challenging as well. Traditionally, pairs are to be found with four methods: correlation method, cointegration method, random spread method, and least distance method. These methods are only focusing on the linear relationship between two stocks. Often, there is no guarantee that the price ratio of such stocks is to be converged in future and failed to explore the complicated relation of multiple stocks. To achieve above goal, Yin Lei researched on a pair trading strategy based on cointegration method for multiple stocks[2] and shows improvement against Buy-and-Hold strategy and Bing Li introduced pairing price prediction based on correlation method[3].

---

[1] (Katanamura et al, 2008)
[2] (Yin & Yu, 2018)
[3] (LI et al., 2019)

This paper will be introducing K-means++ clustering algorithm to break the industrial boundaries of different stocks based on their stock market performances, then apply ElasticNet algorithm trained with AdaBoost weight distributor to mine the underlying relationship within multiple cointegrated stocks hence predict future price movement.

1. Important Theories

1.1 K-means++ Algorithm

K-means clustering algorithm is an unsupervised machine learning clustering algorithm. It will generate K clusters and assigning data points into each one of them while the centroids being adjusted dynamics as new data points being included. The goal is to minimize the Least Squared Error within each cluster based on Euclidean distance

$$E = \sum_{j=1}^{K} \sum_{x \in C_j} ||x - c_j||_2^2 \quad (1)$$

Where Cj is centroids with j=1,2,3…….

The problem with K-means algorithm is that the K initial centroids is assigned randomly, sometimes an outlier will be assigned as initial cluster center which will alter desired outcome. To solve such issue, K-means++ is introduced, when it initializes the centroids, a data point with further distance to the previous assigned center has a higher probability of being assigned as the next centroid.

1.2 ElasticNet Regression

ElasticNet regression was first introduced by Zhen Zhang[4], it is a regression that unites L1 regularization from Lasso and L2 regularization from ridge regression as a penalty

[4] (Zhang et al., 2017)

term to overcome the existing issue of both algorithms. It can offset the effects from both sparsity and non-sparsity data set that are overfitting and low explaining power of features.

ElasticNet Regression can be written as:

$$w = \min_{w} \{ \frac{1}{2N} \sum_{i=1}^{N} ||X * w - y||_2^2 + \lambda * P_\alpha(w) \} \quad (2)$$

where

$$P_\alpha(w) = \frac{1-\alpha}{2} ||w||_2^2 + \alpha ||w||_1 \quad (3)$$

is a penalty term for ElasticNet regression.

λ, α are the non-negative regularization parameter where when $\alpha = 0$, equation (2) becomes a Ridge regression and a Lasso regression when $\alpha = 1$.

1.3 AdaBoost Boosting Algorithm

AdaBoost algorithm is one of the most significant boosting algorithms that was first introduced in 1997 by Freund and Schapire.[5] The core ideology of the algorithm is to learn through multiple iterations on a basic or weak classifier, and on each iteration, the weight of wrongly assigned data from previous classifier will be increased and lower the opposite. The final combination of these weak classifier will result in a strong classifier. These steps will ultimately pay more attention to the wrongly assigned data points with a linear combination of those weak classifier with larger weight on lower classification error learners to construct a strong classifier.

In this paper, the weak classifier is ElasticNet regression where it no longer remains a classification problem, hence modification is needed. The point is to better measure the

[5] (Freund & Schapire, 1999)

residuals so that the weights can be updated through each iteration. AdaBoost regression is converted in following steps.

1. Initialize sample weight

$$W(1) = (w_{1,1}, w_{1,2}, \ldots\ldots, w_{1,m}) \quad (4)$$

where there are m data points $(X_1, y_1), (X_2, y_2), \ldots\ldots, (X_m, y_m)$

$$w_{1,i} = \frac{1}{m} \text{ for } i = 1,2, \ldots\ldots, m$$

2. For k-th iteration:
   a) Use W(k) to train the model to get weak learner $G_W(x)$
   b) Calculate maximum error using $G_k(x)$ on training dataset:

$$\epsilon_k = \max |y_i - G_k(x_i)| \text{ for } i = 1,2, \ldots\ldots, m \quad (5)$$

   c) Calculate relative error for each sample

$$e_{k,i} = \frac{|y_i - G_{k(x_i)}|}{\epsilon_k} \quad (6)$$

   d) Calculate cumulative error

$$e_k = \sum_{i=1}^{m} w_{k,i} e_{k,i} \quad (7)$$

   e) Calculate the coefficient for this learner

$$\alpha_k = \frac{e_k}{1 - e_k} \quad (8)$$

   f) Update W(k+1)

$$w_{k+1,i} = \frac{w_{k,i}}{Z_k} \alpha_k^{1-e_{k,i}} \quad (9)$$

$$where\ Z_k = \sum_{i=1}^{m} w_{k,i} \alpha_k^{1-e_{k,i}}$$

3. Construct strong learner

$$G(x) = \sum_{i=1}^{M} (g(x) * \ln \frac{1}{\alpha_k}) \quad (10)$$

Where g(x) is the medium of all weak learner multiply by its coefficients being learned through the algorithm.

2. Multi-stock pair-trading strategy based on AdaBoost-ElasticNet

2.1 Basic Strategy

Traditional pair-trading strategy is based on statistical support that provides reasonable believe that two historically closely related stock will continue to maintain their price ratio despite short term divergency, and arbitrage can be profitable during this short-term deviation by short on the stock that price increases and long the stock that price decreases. The later convergence will create value from this portfolio.

In this paper, a portfolio strategy is being constructed using following steps. First, SP500 stocks from US stock market is being selected as initial stock pool. SP500 contains stocks with best performance from diversified industries. A K-means++ clustering analysis is done based on their stock market performance measures. Then, one stock is being selected from the clusters which denoted by Y and let other stocks that are highly cointegrated with Y being $X = (x_1, x_2, \ldots, x_n)$. Goal is to find a function $Y = f(X) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$ that can describe the linear combination of those stocks so that a virtual stock Y' can be estimated using this formula for pair-trading with Y. A flowchart follows:

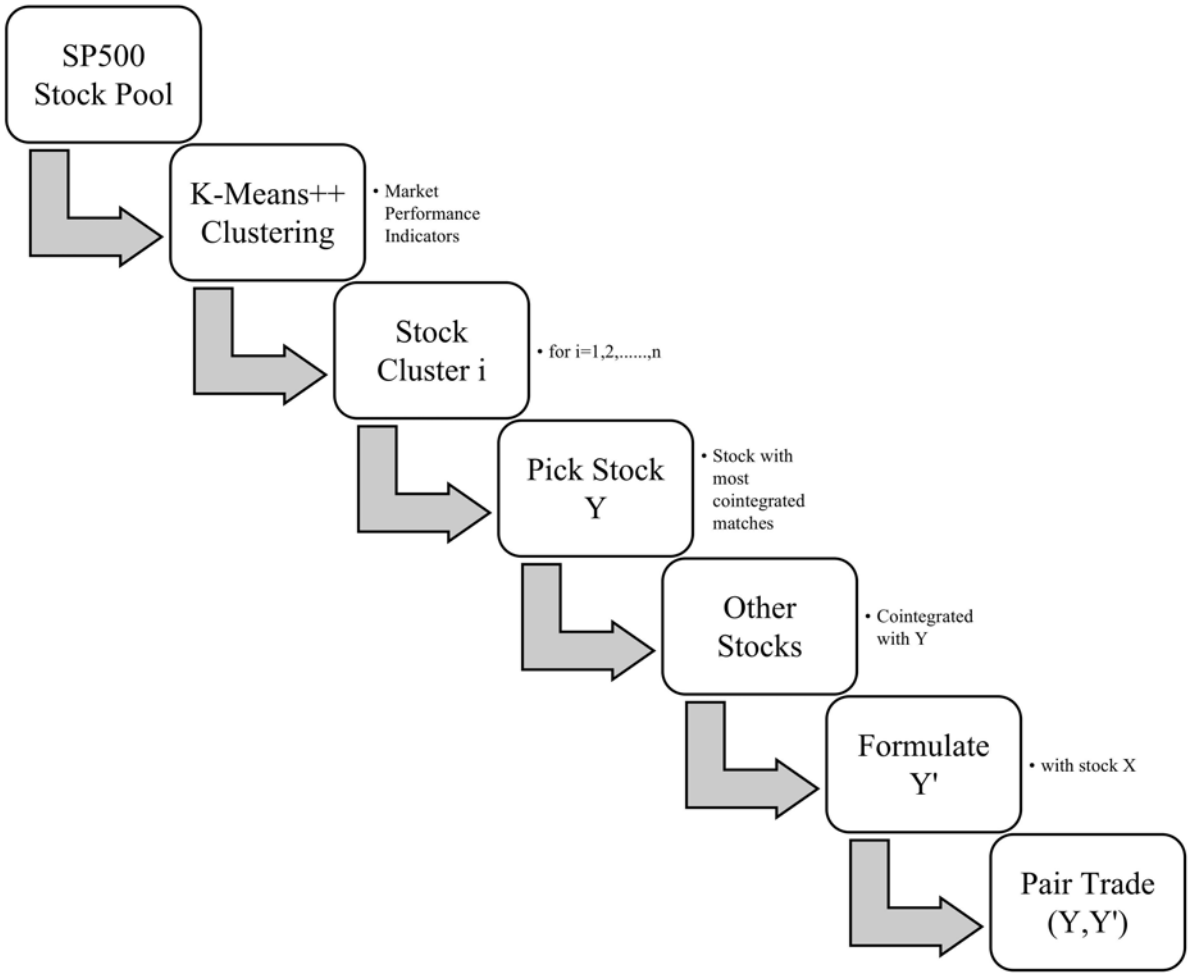

Figure 1 Flowchart of Multi-stock pair-trading strategy based on AdaBoost-ElasticNet

2.2 Data processing

For this experiment, the data was chosen from SP500 Index from Yahoo Finance with start date of 2021-05-19 to end date of 2022-05-19. Data was collected on daily closing prices for each ticker to explore its price series relationships such as trend and momentum. One of the main points using clustering algorithms as a pre-partitioning method is to break any industrial boundaries, hence financial reports are omitted because of two reasons. 1, Same industry may share similar debt ratio, profit measures, may lead to bias clustering. 2,

Financial reports suffers from oddity because the data density is sparse. Therefore, market performance indicators are chosen to represent the profitability and stability of a stock over the year. Indicators are annual average return, annual volatility, and Sharpe ratio. The average is calculated using cumulative total of daily return divided by number of trade days which is 252 in a year. Same method was used for volatility. Sharpe ratio is calculated by annual return divided by annual volatility. Initially, 504 tickers have been recorded, after data cleaning, 501 ticker remains. Data are cleaned in a sense that if a ticker has more than 20% of missing values, it will be removed while other missing values are filled with back-fill method without affecting generosity.

Since every ticker are from different price range, hence a standardization is required for them to be compared on the same level. In this paper, z-score standardization is used

$$x^* = \frac{x - \mu}{\sigma} \quad (11)$$

Where x* is the standardized score, x is the original input, $\mu \; and \; \sigma$ stands for mean and variance of input data.

Since the dimension includes return, volatility and Sharpe ratio, a Principal Component Analysis is needed for better visualization, and due to the fact Sharpe ratio is constructed using return and volatility, PCA is also useful to testify the importance of Sharpe ratio in this model. Here, the principal component is chosen to be 2.

2.3 K-means++ Clustering and Cointegration Analysis

K-means++ clustering is used in this paper. Number of clusters is the only parameter needs tuning. To find the optimized K value for this model, Elbow Method is used. This is one of the most classic methods to find an optimized value K using a plot where y-axis represents distortion and x-axis represents value of K. The tangent line's intersection with the curve is the optimized point.

Cointegration analysis is performed to check for any cointegrated pairs of stocks within each cluster. In this paper, augmented Enger-Granger two-step cointegration test is being used with significance value set to be 0.05. Any p-value greater than 0.05 will not be considered to be cointegrated.

2.4 Trade Methodology

Traditional pair-trading method uses short-long strategy on both stocks to secure margins. However, since virtual stock Y'is constructed using multiple stocks, there is no way of realization on trading such. Hence, a new trading method is proposed following the idea of traditional pair-trading.

Since Y and Y' are highly correlated to each other, it is possible to predict stock Y's movements using virtual stock Y'. Whenever the prediction shows a upward trending is coming, a long position is settled for Y, vice versa.

3. Experiment

3.1 Cluster Analysis

After performing PCA on standardized data for 501 stocks, 95% of information is being used, and the dimension has been reduced to 2-dimension.

Using Elbow Method, the optimized cluster counts for 501 stocks is 6. Hence the number of clusters being settled is 6.

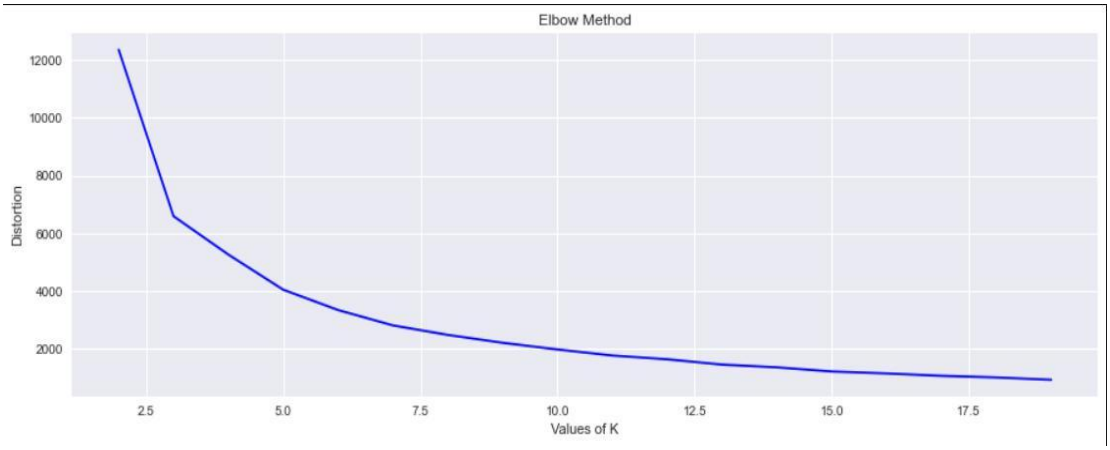

Figure 2. Elbow Method plot for K-means++ Clustering Analysis

Figure 3. Cluster Visualization

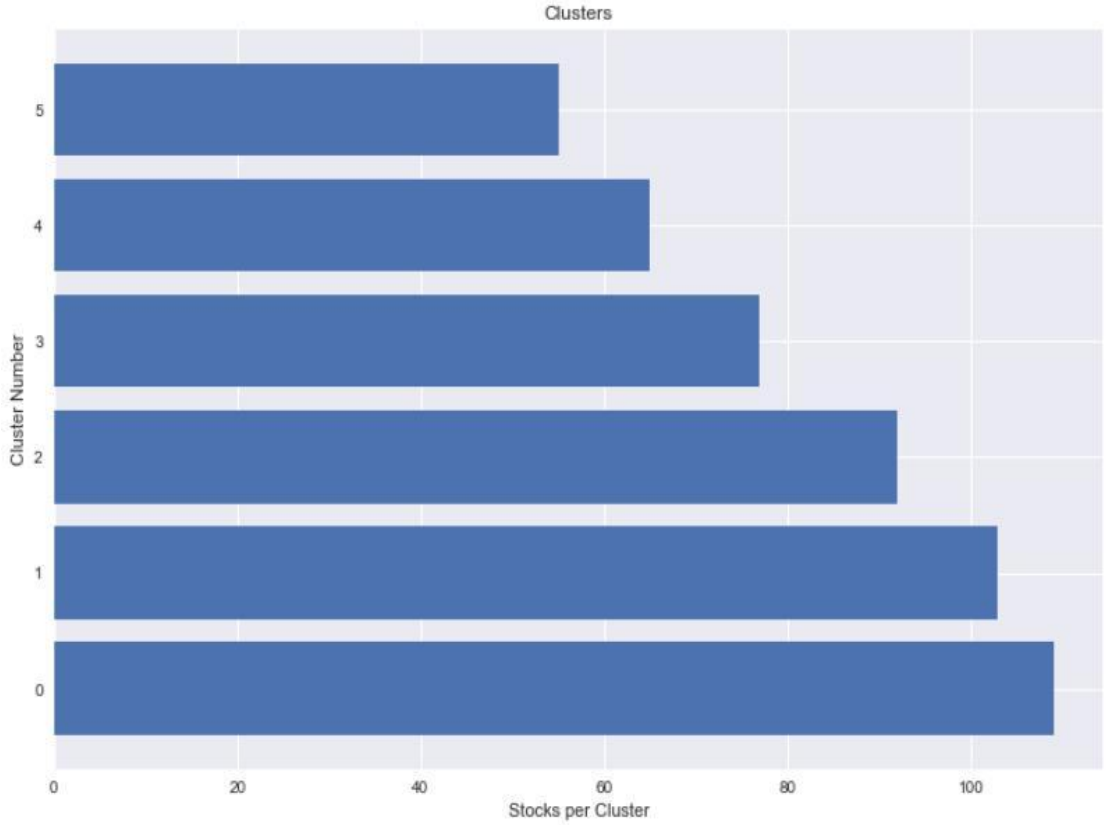

Figure 4. Cluster Content Counts

3.2 Cointegration Test

From the result of Engle-Granger's test, there are 36 pairs of stocks being found in 6 clusters and 41 unique tickers being recorded, hence there exist stocks that are cointegrated with more than one stock.

```
Number of pairs: 36
In those pairs, we found 41 unique tickers.
[('AES', 'MHK'), ('AES', 'SWK'), ('AES', 'WDC'), ('ABMD', 'ALGN'), ('ABMD', 'BIO'), ('ABMD', 'TECH'), ('ABMD', 'KMX'), ('ABMD', 'CARR'), ('ABMD', 'CRL'), ('ABMD', 'CMG'), ('ABMD', 'EFX'), ('ABMD', 'FTV'), ('ABMD', 'GRMN'), ('ABMD', 'IDXX'), ('ABMD', 'ISRG'), ('ABMD', 'MCO'), ('ABMD', 'MSCI'), ('ABMD', 'NDAQ'), ('ABMD', 'NFLX'), ('ABMD', 'PAYC'), ('ABMD', 'SPGI'), ('ABMD', 'CRM'), ('ABMD', 'NOW'), ('ABMD', 'TXN'), ('ABMD', 'UAA'), ('ABMD', 'XYL'), ('AFL', 'HPE'), ('AFL', 'MTB'), ('MMM', 'ABC'), ('MMM', 'WRB'), ('MMM', 'CTRA'), ('MMM', 'GD'), ('ABBV', 'D'), ('ABBV', 'FE'), ('ABBV', 'HSY'), ('ABBV', 'NI')]
```

Figure 5. Names of paired stocks

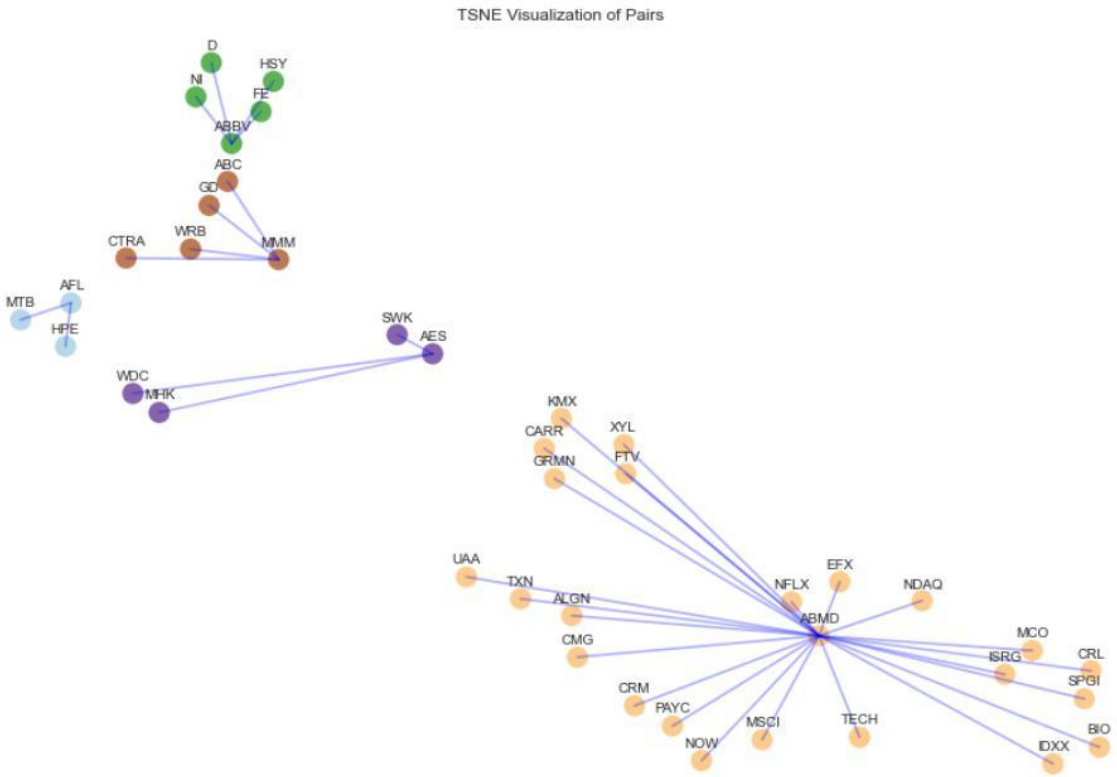

Figure 6. TSNE plot on paired stocks in different clusters.

3.3 Prediction

The main target of proposed portfolio is to ensure that virtual stock Y' constructed with multiple stock X must be closely performing with the actual Y. The closer they are, better margin can be captured. Hence, to measure the performance of the prediction model, several evaluation indicators are being used to make a robust conclusion. The indicators being used are MSE (mean square error), MAPE (mean absolute percentage error) and R2 to measure the explanation power of the regression.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\dot{Y_i} - Y_i)^2 \quad (12)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|\dot{Y_i} - Y_i|}{Y_i}\right) \quad (13)$$

In this experiment, ABMD is chosen to be the target stock and {'GRMN','CARR','KMX','FTV','XYL','NFLX','EFX','NDAQ','MCO','ISRG','CRL','SPGI','BIO','IDXX','TECH','MSCI','NOW','PAYC','CRM','CMG','ALGN','TXN','UAA'} are being used of training data.

To demonstrate that AdaBoost-ElasticNet has meaningful improvement than traditional method, a comparison between OLS, ElasticNet and AdaBoost ElasticNet is being made. OLS is chosen because all tree-based algorithms are not capable of extrapolate.

The prediction period is being settled as '2022-03-19' to '2022-05-19'.

The prediction result is given as follows:

Figure 7. Price Trend Prediction using Real (red), AdaBoost-ElasticNet (blue), ElasticNet (yellow) and OLS (green)

The evaluation is as follows:

	MSE	MAPE	R2
OLS	14.59948	0.044108	0.780256
ElasticNet	13.65929	0.041045	0.807647
AdaBoost	9.897102	0.027466	0.899015

Figure 8. Evaluation matrix of prediction

As demonstrated by above plots, AdaBoost-ElasticNet is have the best performance measures against OLS and ElasticNet. The reason behind is that OLS has trouble with dimensionality and failed to explore the complex relationships within the data. ElasticNet shows some degree of prediction power on the test data yet Boosted algorithm has better performance due to the weight distribution.

3.4 Portfolio Evaluation

[**Methodology**] Portfolio is being constructed using a popular quantitative method in financial industry called simple crossover moving average. It contains three moving averages from the data which are short-term moving average, mid-term moving average and long-term moving average. The idea is, whenever 1. Short-term MA > Mid-term MA and 2. Mid-term MA > Long-term MA, a buy signal is triggered (note: both conditions must be meet at the same time otherwise disregarded), vice versa for sell signal. In this paper, the trading signal is generated using predicted values from Adaboost-ElasticNet prediction, and the comparison is made against real data. In real world trading evaluation, most tests were compared using strategy versus Buy-Hold which means buy and hold the stock until the end of the testing period. Same comparing method is being implemented here. The transaction cost was set to 0.1% of the trade value.

[**Key Outcomes**] The evaluation is made between trade signal created by SCMA from predicted data to real data's Buy-and-Hold strategy. The back-testing result is as follows:

```
Return_Perc: [4, 5, 15] | Sharpe: 6.12518
================================================================================================
SIMPLE PRICE & VOLUME STRATEGY | INSTRUMENT = ABMD | RANGES = [ 4  5 15]
------------------------------------------------------------------------------------------------
PERFORMANCE MEASURES:


Multiple (Strategy):         1.237435
Multiple (Buy-and-Hold):     0.77802
-------------------------------------
Out-/Underperformance:       0.459415


CAGR:                        2.739177
Annualized Mean:             1.350267
Annualized Std:              0.447199
Sharpe Ratio:                6.125184
================================================================================================
```

Figure 9. SCMA backtest result on predicted value

Buying ABMD with 1$ and hold it till the end of the testing period will leave with 0.77802$ where as trade SCMA with 1$ will result in 1.237435$ at the end. The improvement is significant with 59.05% increase to Buy-and-Hold option. Sharpe ratio is 6.12 indicating this is a low-risk portfolio compared to its returns. CAGR is compounded annual growth rate which is another indicator for annual return performance. The calculations are as follows:

$$\text{Sharpe Ratio} = \frac{\text{CAGR}}{\text{Annualized\_standard\_deviation}}$$

$$\text{CAGR} = \left(\frac{\text{End}_{\text{value}}}{\text{Start}_{\text{value}}}\right)^{\frac{1}{\text{number of year}}} - 1$$

$$\text{ASD} = \frac{\text{Standard Deviation of Returns}}{\sqrt{\text{trading days in a year}}}$$

Figure 10. Plot of SCMA vs B-H with predicted signal

[**Extended Comparison**] The eliminated the potential effect of SCMA's outperformance to Buy-and-Hold strategy, a backtest using trade signal created from real data in comparison to Buy-and-Hold option was conducted as well. The result is as follows:

```
=========================================================================================================
SIMPLE PRICE & VOLUME STRATEGY | INSTRUMENT = ABMD | RANGES = [ 4  5 15]
---------------------------------------------------------------------------------------------------------
PERFORMANCE MEASURES:


Multiple (Strategy):         1.109637
Multiple (Buy-and-Hold):     0.77802
-------------------------------------
Out-/Underperformance:       0.331617


CAGR:                        0.904148
Annualized Mean:             0.659369
Annualized Std:              0.396994
Sharpe Ratio:                2.277489
=========================================================================================================
```

Figure 11. SCMA backtest result on real data

The result indicates that under the same strategy (SCMA), the annual multiplier 1.109637 is lower than the predicted value's 1.237435, the improvement on the proposed algorithm along is 11.52%. And Sharpe ratio of 2.28 shows this is riskier than proposed solution.

Figure 12. Plot of SCMA vs B-H with real data's signal

4. Conclusion.

The proposed solution provided in this paper finds underlying pair-trading based arbitrage opportunities with K-Means++ to break industry barriers on stock pool and Adaboost algorithm boosted ElasticNet regression to improve the predictive power of the algorithm. This paper has contributed to a possibly new approach to pair-trading solution on multiple-stocks. Viewing by the result, taking transaction cost in consideration, higher Sharpe ratio and Annual Multiplier indicates that this proposed solution has strong confidence in deploying in real-life situation.

Although great result has been presented, it is vulnerable to market-related policy changes or market trend. If a stock's momentum is being modified by policy change or market trend, the relationship with related stocks may be affected hence reduce the accuracy of prediction. To compensate this weakness, one possible solution is to construct portfolio with multiple pair-tradable stocks from different industries to hedge the risk.

References

Katanamura, T., & Rachev, S, Fabozzi, F. (2008). The Application of Pairs Trading to Energy Futures Markets (PDF). Karlsruhe Institute of Technology. Retrieved 20 January 2015.

Freund, Y., & Schapire, R. E. (1999). *A short introduction to boosting - york university*. Retrieved June 8, 2022, from https://www.yorku.ca/gisweb/eats4400/boost.pdf

LI, B., GAO, B., SUN, J., & YU, C. (2019). An Arbitrage Algorithm Based on AdaBoost-ElasticNet. *Journal of China University of Metrology*, *30*(1).

Yin, L., & Yu, C. (2018). Multi-stock pairs trading method based on cointegration[J]. *Journal of Hubei University (Nature Science)*, *40*(4), 323–326.

Zhang, Z., Lai, Z., & Xu, Y. (2017). Discriminative Elastic-Net Regularized Linear Regression[J]. *IEEE Transactions on Image Processing*, *26*(3), 1466–1481.