NUS INVESTMENT SOCIETY

Quantitative Finance Department

NUS National University of Singapore

27 May 2022

Mean-Variance Modern Portfolio Optimization using Machine Learning

Koh Hong Po[1,3] || Wilfred Edward Tan[1,2] || Wrik Karmakar[1,3] || Sumanth Yalamarty[1,3] || Imraj Singh Sandhu[1,3]

[1] *Quantitative Researcher, Department of Quantitative Finance, NUS Investment Society, Singapore*
[2] *Quantitative Finance, Faculty of Science, National University of Singapore, Singapore*
[3] *Computer Science , Faculty of Computing, National University of Singapore, Singapore*

ABSTRACT

Portfolio Optimization has been of great importance for both professional and retail investors. Successful portfolio construction has allowed investors to manage risk by maintaining a well-diversified portfolio. Mean-variance modern portfolio theory is a mathematical framework that has traditionally aided professional investors to construct a portfolio with minimum risk through the reduction of correlation between assets within a portfolio. The performance of most portfolio optimization methods, including mean-variance modern portfolio theory, has been heavily dependent on the prediction of future stock prices. Although traditional mean-variance portfolio optimization methods work sufficiently to create a minimum risk portfolio, many investors find it hard to achieve their ideal expected returns. This is mainly because of its weak prediction of future stock prices. Our paper uses our PLUS+ algorithm to obtain the least correlated stocks and random forest classifier (RFC) machine learning, with mean-variance optimization in Python programming language. Our portfolio optimizer has allowed better prediction of future stock prices and significant reduction in asset correlation specific risk. As a result, our optimized portfolio has outperformed the Standard and Poor's 500 (S&P 500) index and traditional mean-variance optimization from a return, risk, and return-risk ratio perspective.

1. Introduction

Portfolio optimization is the construction of a portfolio consisting of the best selection of financial assets. Portfolio optimization has been an important investment and trading strategy utilized by both retail and institutional investors to generate high returns within a reasonable margin of risk. Generally, optimized portfolios are designed to maximize expected returns while minimizing risks [1].

Mean-Variance Analysis is a modern portfolio optimization theory derived from a mathematical framework which prioritizes the less risky portfolio assuming that expected returns are the same. Portfolios with increased risks are only chosen when there are higher expected returns to justify the risk to reward [2]. Risk can be primarily categorized into systematic risk and specific risk. Specific risk refers to the specific risk associated with that particular asset while systematic risk refers to the market risk common to all assets. Specific risk can be mitigated through the reduction of correlation between assets within a portfolio by means of diversification. From a mathematical theoretical perspective, as the number of uncorrelated assets approach the limit infinity, specific risk approaches the zero. Systematic risk, on the other hand, cannot be reduced by diversification. Systematic risk can be represented by the standard deviation of the portfolio and mitigated through simultaneously using long and short strategies to create a market neutral portfolio. Therefore, Mean-Variance portfolio optimization factors the assets' expected returns, volatility, and correlation with other assets in the portfolio. [1, 3, 4]

Expected return refers to the future performance of the stock. It is a prediction of the future direction and amplitude which the stock will take. Accuracy of the expected return forecast is crucial towards the performance of the Mean-Variance Optimized portfolio. Existing studies have relied on statistical and machine learning methods to predict expected returns.

Common time series statistical methods include autoregressive conditional heteroscedasticity (ARCH) [5], generalized autoregressive conditional heteroscedasticity (GARCH) [6], and autoregressive integrated moving average (ARIMA) [7] models used to forecast expected returns by analyzing historical pricing. Common machine learning models include support vector regression (SVR) [8, 9], logistic regression (LR) [10, 11], and random forest classifier (RFC) [12, 13]. Previous studies have indicated that machine learning is better suited for prediction of non-linear and non-stationary assets like stock prices [14].

Volatility, representing systematic risk can be quantified using historical standard deviation of the asset. Alternatively, specific risk will be mitigated by constructing a portfolio with the least correlated assets [15].
In Summary, our paper proposes to design an optimized portfolio using Mean-Variance portfolio optimization theory and machine learning to predict an asset's expected return. Our approach can be summarized as the following:

1. Portfolio Uncorrelated Stock (PLUS⁺): We designed an algorithm named PLUS⁺ to choose n least correlated assets to form a portfolio from a universe on n assets.
(n refers to a numerical constant variable which can be adjusted to the number of assets desired in a portfolio and universe refers to all assets or a subset of assets.)

2. Compare accuracy of machine learning models in prediction of future expected returns.

3. Generate expected returns of n assets chosen by PLUS⁺ using most accurate machine learning model.

4. Using Python programming language to computerize the weightage of each asset in our PLUS⁺ portfolio.

2. Methodology

This section introduces the methods used in our approach.

2.1. Data Used

For this paper, our team focused only on stocks within the Standard and Poor's 500 (S&P 500) which we define to be our universe of stocks. We set the number of stocks in our portfolio for this paper to n = 10. Python support libraries NumPy and pandas were used for the reading, cleaning, and manipulation of data. Historical price data for our universe of stocks was scrapped from Yahoo Finance using the yfinance API for python. Stock price data outside 2 years of training data was scrapped outside of the Coronavirus disease (COVID-19) period, to simulate normal market conditions. The model used 2 years of historical price as training data from 1st January 2018 to 31st December 2019, to predict future stock price movements from 1st November 2020 to 1st February 2021 (3-months), and 1st November 2020 to 1st November 2021 (1 year).

2.2. Portfolio Uncorrelated Stock (PLUS⁺)

Our goal in this paper is to find the "least correlated portfolio" defined by minimizing the total pairwise correlation between all the stocks in the portfolio.

The problem can be modelled using a graph, setting the nodes to be the stocks and the undirected edges to be the pairwise correlation between the two stocks (nodes) it is joining.

A graph theory visualization modelling the problem is shown below:

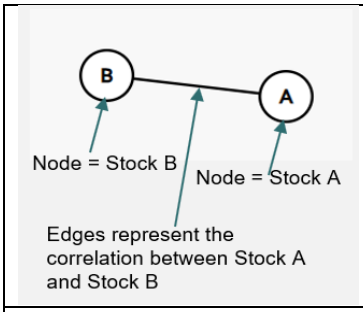

Figure 2.2A: Representative graph theory diagram on a universe of k = 2 stocks. (Each node represents a specific stock and edges represent correlation between 2 respective stocks.)

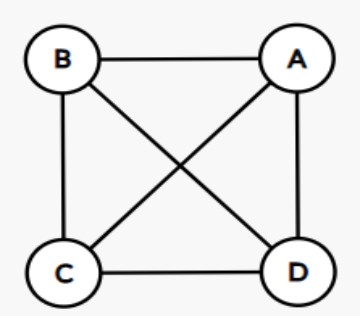

Figure 2.2B: Representative graph theory diagram on a universe of k = 4 stocks. (Each node represents a specific stock and edges represent correlation between 2 respective stocks.)

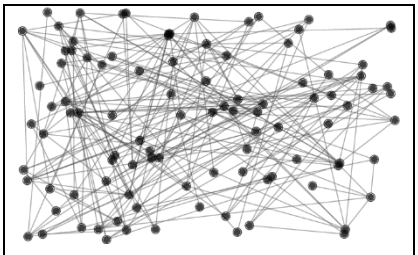

Figure 2.2C: Representative graph theory diagram on a universe of 100 stocks. (Each node represents a specific stock and edges represent correlation between 2 respective stocks.)

The canonical problem that arises is thus looking for the subset of k nodes which minimizes all the edges between the k nodes. This problem is NP-Hard through the following reduction from the NP-Complete k-clique [16] problem for arbitrary k.

1. Create a new weighted graph G'.
2. For each edge in G, create the same edge in G' with weight -1.
3. Find the subgraph H of n nodes which minimizes all the edges between the k nodes from G'
4. H will thus contain (k-1)! edges with total weight of -(k-1)! if H is fully connected. If H has less than -(k-1)! total weight, a k-sized clique does not exist in G'
5. If H has total weight of -(k-1), the corresponding subgraph in G will be a k-sized clique.

For a fixed k, it follows much like the k-clique problem for fixed k, inspecting each k-sized combination would result in an $O(n^k)$ algorithm, exponential in k.

Therefore, choosing the S&P 500 as our universe and a portfolio of n = 10 stocks will result in a combination of 500 choose 10, which is equivalent to $2.46 * 10^2$ combinations. As a result, our team took a greedy approach to create a 'good enough' k-sized subgraph to filter the least correlated stocks based on the following criteria:

a) node that has most numbers of edges with correlation less than a threshold **[Figure 2.2D]**

b) node where the sum of the edges is the lowest **[Figure 2.2E]**

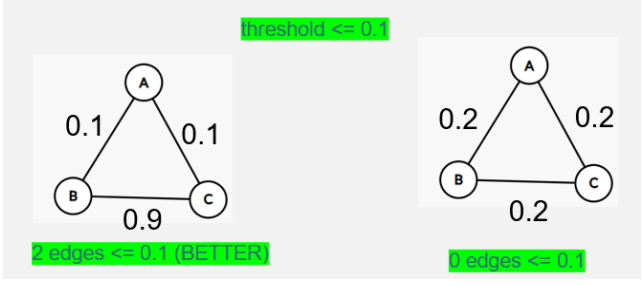

Figure 2.2D: Representative graph theory diagram on a subgraph of k = 3 stocks using criteria a. (Each node represents a specific stock and edges represent correlation between 2 respective stocks.)

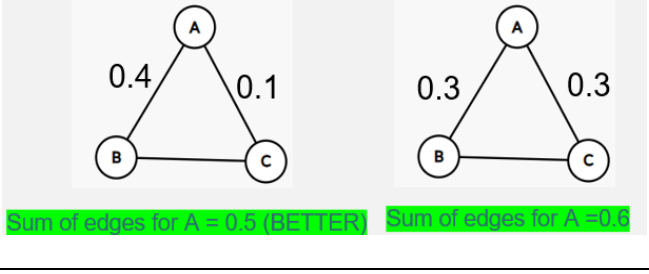

Figure 2.2E: Representative graph theory diagram on a subgraph of k = 3 stocks using criteria b. (Each node represents a specific stock and edges represent correlation between 2 respective stocks.)

The initial k sized subgraph was further improved by the following algorithm:

Algorithm 1: PLUS

```
Result: Write here the result
currChosenStocks = getInitialChosenStocks(allStocks);
currUnchosenStocks = allStocks - currChosenStocks;
currChosenStocksCorr = getPorfolioCorr(currChosenStocks);
for i = 1...len(allStocks) do
    toAdd = Null;
    toRemove = Null;
    minCorrDiff = 0;
    for unchosenStock in currUnchosenStocks do
        for unchosenStock in currUnchosenStocks do
            if getPorfolioCorr(currChosenStocks - chosenStock + unchosenStock) -
              currChosenStocksCorr ≤ minCorrDiff then
                toAdd = unchosenStock;
                toRemove = chosenStock;
            else
            end
        end
    end
    if toAdd != Null then
        currChosenStocks - toRemove + toAdd;
        currUnchosenStocks + toRemove - toAdd;
        currChosenStocksCorr = getPorfolioCorr(currChosenStocks);
    else
        return currChosenStocks
    end
    return currChosenStocks;
end
```

The chosen set of stock is improved by iterating every possible replacement of a chosen stock and an unchosen stock and finding the replacement which decreases the total correlation of the set the most. The maximum iterations performed is equals to the number of stocks in the universe.

Time complexity is as follow:

T(n) = #iteration of outermost for loop * T(outermost for loop)

= n * #replacement pairs * O(1) + O(1)

= $O(n^2k)$

Where n = the number of stocks in the universe and k = the number of stocks in the portfolio.

In our testing, the number of iterations does not go past #stocks/10 iterations.

2.3. Forecasting stock prices using Machine Learning algorithm

Classification machine learning model were used to forecast whether the price of a stock increases in the next time step (30, 90, 180, 365 days). Then the expected return is calculated by taking the expected value over the Half-Normal distribution with the simple mean and standard deviation of the last time step.

The following technical analysis indicators were used as input/ feature vectors for the machine learning models: Simple Moving Average, Average True Range, +/- Directional Movement Index, +/- optional directional index and Moving Average Convergence Divergence (MACD).

The models explored were Logistic Regression (LR), Linear/Radial/Poly Support Vector Machines (SVMs), AdaBoost-Decision-Tree-Classifier, and Random Forest Classifier (RFC) machine learning and their accuracy was measured against actual returns. Implementation of these models were from Python machine learning libraries scikit-learn and TensorFlow

2.4. Using Python programming language to computerize the weightage of each asset in our PLUS+ portfolio using Mean-Variance analysis

According to Mean-Variance optimization theory, asset specific risk can be reduced through the reduction of correlation between assets in the portfolio to find the Global Minimum Variance Portfolio (GMVP). In theory, as the number of uncorrelated assets in a portfolio approaches infinity, the portfolio risk decreases and approaches zero. Our team implemented the Mean-Variance Analysis to a portfolio on the n = 10 chosen stock by PLUS+ using the PyPortfolioOpt extension library in Python. **[Figure 2.4A]**

Rate of Return of Asset:

$$r = \frac{W_1 - W_0}{W_0} = \frac{W_1}{W_0} - 1.$$

(Current amount – initial amount) / initial amount

Risk of Asset = Standard Deviation of the rate of return of asset

$$\sigma_i = \sqrt{\mathrm{Var}(r_i)}$$

Measure of association between two assets, correlation co-efficient (rho)

$$\rho_{i,j} = \frac{\sigma_{i,j}}{\sigma_i \sigma_j}$$

Portfolio Mean (Expected Value of Portfolio's Rate of return):

$$\mu_p = \mathbf{E}(r_p) = \sum_{i=1}^{n} w_i \mu_i \qquad r_p = \sum_{i=1}^{n} w_i r_i$$

R_p: Portfolio Rate of Return
Sum of (weightage * returns) of stocks

Portfolio Variance

$$\sigma_p^2 = \mathrm{Var}(r_p) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_{i,j}$$

Variance Risk can be reduced as uncorrelated assets owned tends to infinity:

However, even if you own infinite uncorrelated assets, systematic market-wide risk cannot be eliminated:

$$\bar{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \sigma_i^2 \text{ and } \bar{\phi} = \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \sigma_{i,j}.$$

$$\text{Suppose that } \frac{1}{n} \sum_{i=1}^{n} \sigma_i^2 \to \sigma^2 \text{ and } \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \sigma_{i,j} \to \phi \text{ as } n \to \infty.$$

$$\sigma_p^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2 + \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \sigma_{i,j} \to \phi$$

Finding Global Minimum-variance portfolio with individual weights for each stock in portfolio

Portfolio of two assets:

$$\mu_p = \alpha \mu_1 + (1 - \alpha) \mu_2$$

$$\sigma_p^2 = \alpha^2 \sigma_1^2 + (1-\alpha)^2 \sigma_2^2 + 2\alpha(1-\alpha)\sigma_{1,2} = \alpha^2 \sigma_1^2 + (1-\alpha)^2 \sigma_2^2 + 2\alpha(1-\alpha)\rho_{1,2}\sigma_1\sigma_2$$

Minimum Portfolio of two assets:

$$\alpha = \alpha^* = \frac{\sigma_2(\sigma_2 - \rho_{1,2}\sigma_1)}{\sigma_1^2 + \sigma_2^2 - 2\rho_{1,2}\sigma_1\sigma_2}$$

$$(\sigma_p^2)^* = \frac{\sigma_1^2 \sigma_2^2 (1 - \rho_{1,2}^2)}{\sigma_1^2 + \sigma_2^2 - 2\rho_{1,2}\sigma_1\sigma_2}$$

$$(\mu_p)^* = \alpha^* \mu_1 + (1 - \alpha^*) \mu_2$$

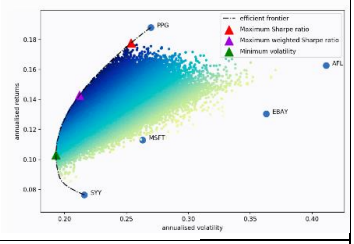

Figure 2.4A: Representative mathematical framework of Mean-Variance Analysis.

3. Results and Discussion

This section discusses our results.

3.1. Portfolio Uncorrelated Stock (PLUS+) reduces average correlation between stocks within portfolio from 0.76 to 0.12

Our PLUS+ algorithm reduced the pairwise correlation of the stocks in a portfolio of n = 10 stocks from an initial average of 0.76 to 0.12 in polynomial time ($O(n^2k)$).

The best results for the initial set were yielded by finding nodes that has the most numbers of edges with correlation less than a threshold with a threshold of 0.5.

PLUS+ allows us to narrow down a universe of stocks to select the 'best', least correlated subset of stocks to construct a diversified portfolio. This significantly reduces asset specific risk on top of the traditional mean variance analysis to minimize portfolio risk.

3.2. Random Forest Classifier (RFC) outperforms other machine learning models for expected returns forecast accuracy

Model	Average accuracy
Random Forest Classification	0.8932
Support Vector Machine (Linear)	0.88465
Logistic regression	0.88312
Adaboost	0.85785
Support Vector Machine (Poly)	0.80785
Support Vector Machine (Radial)	0.70812

Figure 3.2A: Representative python output of average accuracy of machine learning models for predicting future expected returns

In terms of the predicting future expected returns, Random Forest Classifier (RFC) had the greatest accuracy compared to Logistic Regression (LR), Linear/Radial/Poly Support Vector Machines (SVMs), and AdaBoost-Decision-Tree-Classifier machine learning models.

As a result, our team decided to focus on Random Forest Classifier (RFC) machine learning model for predicting the expected returns for our portfolio optimization.

3.3. Results of our Mean-Variance portfolio optimization with Random Forest Classifier (RFC) and PLUS+ against standard benchmarks

```
Expected annual return: 25.4%
Annual volatility: 14.8%
Sharpe Ratio: 1.58
```

Figure 3.3A: Representative python output of annual return, annual volatility, and Sharpe ratio of our portfolio optimization using:
1) Portfolio Uncorrelated Stock (PLUS+)
2) Random Forest Classifier (RFC)
3) Mean-Variance Analysis

```
Expected annual return: 14.0%
Annual volatility: 12.4%
Sharpe Ratio: 0.97
```

Figure 3.3B: Representative python output of annual return, annual volatility, and Sharpe ratio of the traditional portfolio optimization using historical mean for expected returns as a comparison benchmark

Our optimized portfolio using Portfolio Uncorrelated Stock (PLUS+), Random Forest Classifier (RFC), and Mean-Variance Analysis generated a higher expected annual return of 25.4 % in comparison to 14% when using a traditional mean-variance analysis with historical mean. A return of 10% after accounting for transaction fee between 0.5% to 1% is considered a relatively good return on investment by investors. As our model refactors only every 90 days, and focusses more on an investment horizon of 3 months to a year, transaction cost are minute and negligible due to stability of weights. Our portfolio also outperformed the traditional mean variance analysis with a Sharpe ratio of 1.58 to 0.97. A Sharpe ratio above 1 is considered adequate while a Sharpe ratio above 1.5 is considered good. Furthermore, our portfolio has volatility percentage on 14.8 % which is within an acceptable range. Volatility generally fluctuates between 10% and 20%, averaging around 15%. Despite our portfolio having a higher volatility by 2.4% compared to the traditional mean-variance analysis, our higher annual expected annual return justifies the risk to reward. **[Figure 3.3A and Figure 3.3B]**

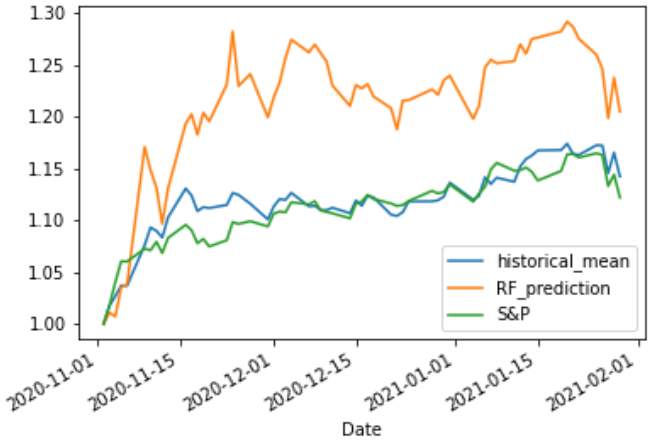

Figure 3.3C: Representative python line graph of our portfolio against benchmarks for a period of 3 months from November 2020 to February 2021.

Orange Line: Our portfolio optimized portfolio using Portfolio Uncorrelated Stock (PLUS+), Random Forest Classifier (RFC), and Mean-Variance Analysis.

Blue Line: Traditional portfolio optimization by Mean Variance Analysis using historical mean. *(Benchmark)*

Green Line: Standard and Poor's 500 (S&P 500) *(Benchmark)*

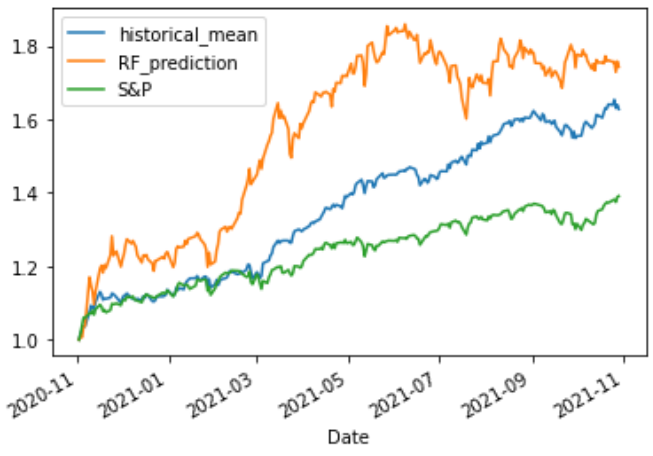

Figure 3.3D: Representative python line graph of our portfolio against benchmarks for a period of 1 year from November 2020 to 2021.

Orange Line: Our portfolio optimized portfolio using Portfolio Uncorrelated Stock (PLUS+), Random Forest Classifier (RFC), and Mean-Variance Analysis.

Blue Line: Traditional portfolio optimization by Mean Variance Analysis using historical mean. *(Benchmark)*

Green Line: Standard and Poor's 500 index (S&P 500) *(Benchmark)*

From a portfolio performance perspective, our optimized portfolio using Portfolio Uncorrelated Stock (PLUS+), Random Forest Classifier (RFC), and Mean-Variance Analysis outperformed both the Standard and Poor's 500 (S&P 500) index and traditional portfolio optimization by Mean Variance Analysis using historical mean benchmarks for the 3 months and 1 year time horizon. 3 months and 1 year is a good time horizon for larger investment firms managing large capital where constant refactoring will incur unjustifiably large transaction costs. Our portfolio optimizer can also be comfortably used by retail swing traders to minimize their risk while generating reasonable returns.

4. Conclusion

This paper proposes a machine learning approach towards constructing a portfolio using mean-variance portfolio optimization. Our Portfolio optimizer is built by mainly 3 parts: Portfolio Uncorrelated Stock (PLUS+), Random Forest Classifier (RFC), and Mean-Variance Analysis. Our Portfolio's refactoring time horizon is ideally 3 months and caters to retail swing traders and larger firms' importance of weight stability due to transaction fees. Our model has proven to outperform the S&P500 index and traditional mean variance analysis for portfolio optimization. Machine Learning methods such as our Random Forest Classifier (RFC) approach allows a more accurate forecast of expected returns required for the mean-variance analysis, allowing the mathematical framework of mean variance analysis to have a better weight allocation.

In summary, mean variance analysis combined with PLUS+ and RFC machine learning outperforms benchmarks such as the S&P500 index and traditional mean variance optimization in terms of returns, Sharpe ratio, and return-risk ratio.

5. Future works

Although this research paper provides valuable insights on the importance of applying computing technology of machine learning and algorithms to traditional portfolio optimization frameworks, our paper research lacks to address systematic crash risk which can be mitigated through hedging. A possible direction for our team would be to look into multiple asset classes to hedge against severe market losses [17]. Natural language processing can also be introduced towards allowing our machines to better understand consumer and market sentiment around the world, and to analyze financial and credit statements for a more fundamental and qualitative perspective.

References:

[1] T. Bodnar, S. Mazur, Y. Okhrin, Bayesian estimation of the global minimum variance portfolio, European J. Oper. Res. 256 (1) (2017) 292–307.

[2] Feldstein, M. S. (1969). Mean-variance analysis in the theory of liquidity preference and portfolio selection. The Review of Economic Studies, 36(1), 5-12.

[3] Renn O, Klinke A. Systemic risks: a new challenge for risk management. EMBO Rep. 2004;5 Spec No(Suppl 1):S41-S46. doi:10.1038/sj.embor.7400227

[4] Delpini, D., Battiston, S., Caldarelli, G., & Riccaboni, M. (2019). Systemic risk from investment similarities. PLoS One, 14(5), e0217141.

[5] R. Engle, Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, Econometrica 50 (4) (1982) 987–1007.

[6] T. Bollerslev, Generalized autoregressive conditional heteroskedasticity, J.Econom. 31 (3) (1986) 307–327

[7] G. Box, G. Jenkins, Time series analysis: forecasting and control, J. Time 31 (4) (1976) 238–242.

[8] W.C. Hong, M.W. Li, J. Geng, Y. Zhang, Novel chaotic bat algorithm for forecasting complex motion of floating platforms, Appl. Math. Model. 72 (2019) 425–443

[9] R. Chen, C.Y. Liang, W.C. Hong, D.X. Gu, Forecasting holiday daily touristflow based on seasonal support vector regression with adaptive genetic algorithm, Appl. Soft. Comput. 26 (2015) 435–443.

[10] Wright, R. E. (1995). Logistic regression.

[11] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

[12] Thakur, M., & Kumar, D. (2018). A hybrid financial trading support system using multi-category classifiers and random forest. Applied Soft Computing, 67, 337-349.

[13] Liu, Y., Wang, Y., & Zhang, J. (2012, September). New machine learning algorithm: Random forest. In International Conference on Information Computing and Applications (pp. 246-252). Springer, Berlin, Heidelberg.

[14] W. Wang, W. Li, N. Zhang, K. Liu, Portfolio formation with preselectionusing deep learning from long-term financial data, Expert Syst. Appl. 143 (2020) 113042.

[15] Markowitz, H. M., & Todd, G. P. (2000). Mean-variance analysis in portfolio choice and capital markets (Vol. 66). John Wiley & Sons.

[16] Chen, Jianer; Huang, Xiuzhen; Kanj, Iyad A.; Xia, Ge (2006), "Strong computational lower bounds via parameterized complexity", Journal of Computer and System Sciences, 72 (8): 1346–1367.

[17] Zhu, S., Zhu, W., Pei, X., & Cui, X. (2020). Hedging crash risk in optimal portfolio selection. Journal of Banking & Finance, 119, 105905.

Disclaimer

This research material has been prepared by NUS Invest. NUS Invest specifically prohibits the redistribution of this material in whole or in part without the written permission of NUS Invest. The research officer(s) primarily responsible for the content of this research material, in whole or in part, certifies that their views are accurately expressed, and they will not receive direct or indirect compensation in exchange for expressing specific recommendations or views in this research material. Whilst we have taken all reasonable care to ensure that the information contained in this publication is not untrue or misleading at the time of publication, we cannot guarantee its accuracy or completeness, and you should not act on it without first independently verifying its contents. Any opinion or estimate contained in this report is subject to change without notice. We have not given any consideration to and we have not made any investigation of the investment objectives, financial situation or particular needs of the recipient or any class of persons, and accordingly, no warranty whatsoever is given and no liability whatsoever is accepted for any loss arising whether directly or indirectly as a result of the recipient or any class of persons acting on such information or opinion or estimate. You may wish to seek advice from a financial adviser regarding the suitability of the securities mentioned herein, taking into consideration your investment objectives, financial situation or particular needs, before making a commitment to invest in the securities. This report is published solely for information purposes, it does not constitute an advertisement and is not to be construed as a solicitation or an offer to buy or sell any securities or related financial instruments. No representation or warranty, either expressed or implied, is provided in relation to the accuracy, completeness or reliability of the information contained herein. The research material should not be regarded by recipients as a substitute for the exercise of their own judgement. Any opinions expressed in this research material are subject to change without notice.