Portfolio Construction with K-Means Clustering Algorithm Based on Three Factors

Bilal Aslam[1,*], Rubaiyat Ahsan Bhuiyan[1], and Changyong Zhang[1]
Department of Accounting, Finance and Economics, Faculty of Business, Curtin University, Miri, Malaysia

*Corresponding author: bilal@postgrad.curtin.edu.my

Abstract. Constructing a portfolio from a large number of active stocks is a critical as well as challenging investment decision due to high volatility and biased decision making. The abundance and availability of _nancial data gives machine learning (ML) an advantage to optimize investment decisions. The k-means algorithm is used to cluster observations into di_erent groups, each of which contains those with similar properties. In this paper, three factors are considered to cluster stocks and select clusters with best performing stocks for portfolio construction. It enhances the cardinal investment decision of stock selection to construct optimized portfolios. The out-of-sample performance demonstrates high economic gains from the proposed strategy with an average Sharpe ratio of 0.7.

Keywords: Clustering, k-means algorithm, machine learning, portfolio construction, stock selection

1. Introduction

On average, investors receive higher returns by investing in the stock market, the return from which tends to be well in excess of the inflation [Brealey et al., 2020]. In the meantime, the time series of the historical prices of the stocks and indices indicate that the stock market is highly volatile, while different biases and illogical decision making also affect stock selection. Investors tend to overweight the events that are unlikely to happen and the bad news than good news. As a result, they become more cautious about losses rather than gains [Barberis, 2013; Rozin and Royzman, 2001], although more recent studies suggest that they should overweight stocks that have positive news and underweight stocks with negative news [Cao et al., 2017]. Thus, because of randomness in the market together with biases in human behavior related to stock selection, managing risk-adjusted portfolios is a difficult practice. A quantitative and fixed strategy is hence necessary to select stocks for constructing risk-adjusted equity portfolios.

In stock portfolio construction, the expected return of individual stocks is critical and a combination of stocks generally reduces the portfolio risk [Guerard et al., 2020]. The prime concentration of investment decision makers is therefore on analysing and screening stocks that can outperform the market [Tan et al.,2019]. The possible gain of even modest improvement in the ability to select well-performing stocks canbe potentially significant [Mondal et al., 2019]. One prominent strategy to sort out promising stocks is the stock price momentum, which is persistent across different asset classes and markets [Wiest, 2022]. It is a phenomenon in which stocks that have recently enjoyed high returns are likely to experience better returns in the future, indicating the existence of return continuation [Gupta and Kelly, 2019].

The stock market normally contains a relatively large number of stocks, the dynamics of whose prices is affected by various variables. Here comes in machine learning (ML), which is concerned with employing big data to learn the relationships among variables, make predictions and pattern recognition. In stock

investment, the availability of large data gives machine learning an advantage over traditional approaches. Consequently, ML techniques are becoming increasingly important and inevitable tools in the finance sector and are expected to increasingly dominate the world of economics and finance [Hull, 2021; Ndikum, 2020]. One popular ML technique to recognize patterns in data is the *k*-means algorithm that divides data into different clusters.

The algorithm has been intensely applied to cluster stocks for investment, for example, by grouping highly correlated securities for performing mean-variance portfolio optimization [Ren, 2005]. In additionto correlation networks to ascertain the correlation structure of financial assets, average of financial ratios including the revenues to assets and net-income to assets ratios are used to cluster stocks [Marvin, 2015],as well as the continuous trend characteristics of the market to construct portfolios or clustering to groupthe least risky stocks [Soleymani and Vasighi, 2022; Wu et al., 2022].

Taking a different approach for investment based on momentum strategy, this paper employs the *k*-means algorithm to identify promising stocks for creating risk-adjusted portfolios that can outperform the market and offer a higher Sharpe ratio.

The rest of the paper is organized as follows. In Section 2, the *k*-means algorithm is presented, as well as the three clustering factors. In Section 3, the algorithm is employed and the portfolios are constructed followed by an analysis on portfolio performance. Section 4 concludes the study.

2. Algorithm and clustering factors

In this section, the k-means clustering algorithm is briefly introduced in Section 2.1 and the three clustering factors are discussed in Section 2.2.

2.1. K-Means clustering

As an unsupervised machine learning technique used for data classification, the *k*-means clustering categorizes individual elements into clusters based on a distance or dissimilarity function [Klaas, 2019]. Elements in a cluster have similar properties and are distinct from those in other clusters. To cluster observations,a distance measure is needed. When there are only two features, x and y, the observations can be plottedon a two-dimensional grid. For two observations A and B, suppose that for observation A, $x = x_A$ and$y = y_A$, and for observation B, $x = x_B$ and $y = y_B$. The Euclidean distance between A and B is given by

$$d(A,B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

This distance measure can easily be extended to more observations with large features. Assume that there are observations on m features and the value of the jth feature for the ith observation is v_{ij}. The distance between the pth observation and the qth observation is

$$d(p,q) = \sqrt{\sum_{j=1}^{m} (v_{pj} - v_{qj})^2}$$

Another important concept in the *k*-means algorithm is the center of a cluster. Suppose that a certainset of observations is considered as a cluster. The center is then computed by averaging the values of allfeatures for each of the observations in the cluster. The *k*-means algorithm minimizes inertia which is thetotal sum of squared distances of the observations from their cluster centers. One popular approach to select the optimal number of clusters k is the elbow method. It involves continuing to increase k until theimprovement in the inertia is relatively small. The *k*-means algorithm hence works as follows [Hull, 2021]:

1. Choose k random points as cluster centers.
2. Assign each observation to the nearest cluster center.
3. Calculate new cluster centers.
4. If cluster centers are changed, go to step 2; Otherwise, stop.

2.2. Clustering factors

A rational investor seeks a better risk-return ratio and stable stocks with higher returns. In applying the k-means algorithm technique to identify and cluster potential promising stocks for investment, three factors are hence considered, μ, μ/σ, and AAT, where μ is the annual return of the stock, σ is the standard deviation of the return, and AAT = ¼($2A_T + S_T$), with S_T being the final price and A_T being the average price of the stock during the estimation time window from t =0 to t = T.

To further reduce the underlying volatility in the stock price in the field of derivatives, a path-dependent option payoff is introduced [Aslam and Zhang, 2022]. The payoff function of the so-called average Asian call option is given by ¼max($2A_T + S_T - 3K$, 0) where K is the strike price. The factor AAT, as a combination of the final price S_T when t = T and the average price A_T during the time period from t = 0 to t = T, is deduced from this option payoff to select relatively stable stocks that perform better during the time period used to cluster stocks into different groups.

3. Portfolio construction and performance

In this section, the sample data are first described in Section 3.1. The performance of the constructed portfolios is then presented in Section 3.2 and Section 3.3.

3.1. Data

The sample data include all active stocks in the Malaysian stock market from 2010 to 2021. There are 695 such stocks and the analysis period is twelve years. The daily stock prices are collected from Yahoo Finance. The time window considered covers both the bull market when the KLSE index increased from 1293 on 3 January 2010 to 1886 on 15 June 2014 and the bear market when the index declined from 1888 on 15 April 2018 to 1568 on 31 December 2021. The twelve-year performance from 2010 to 2021 of KLSE index is given in Figure 1. The stocks selected to form portfolios in Section 3.2 and Section 3.3 are as listed in Figure 2.

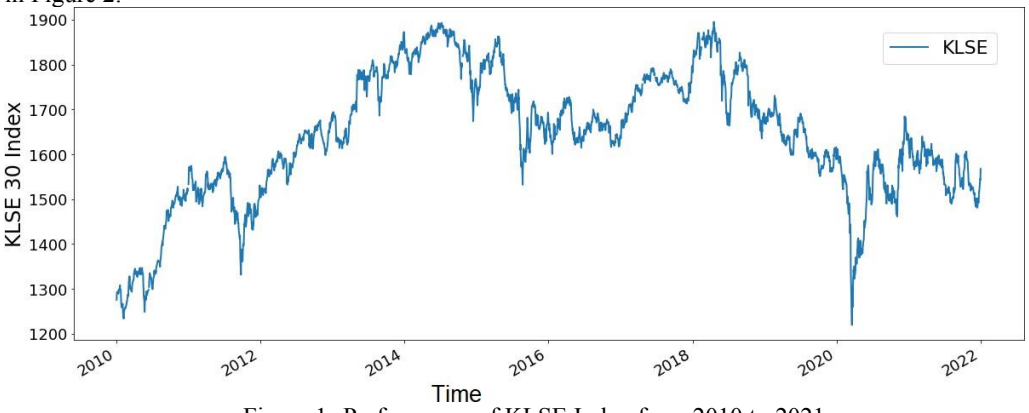

Figure 1: Performance of KLSE Index from 2010 to 2021

A: 19 stocks based on 3-year estimation window.									
0049.KL	0060.KL	0075.KL	0107.KL	0113.KL	0120.KL	0146.KL	4847.KL	7003.KL	7007.KL
7105.KL	7412.KL	9695.KL	0022.KL	5040.KL	5068.KL	5102.KL	9792.KL	9938.KL	
B: 29 stocks based on 5-year estimation window.									
0078.KL	5068.KL	7105.KL	7293.KL	8206.KL	0023.KL	0024.KL	0097.KL	0113.KL	0117.KL
0120.KL	0138.KL	2852.KL	3034.KL	5040.KL	5062.KL	5102.KL	6139.KL	6637.KL	7003.KL
7155.KL	7251.KL	7412.KL	8397.KL	8966.KL	9539.KL	9598.KL	9792.KL	9938.KL	
C: 21 stocks selected before COVID-19 pandemic.									
0097.KL	0127.KL	7105.KL	7160.KL	7293.KL	0078.KL	0113.KL	0120.KL	0138.KL	3034.KL
4731.KL	5068.KL	5102.KL	6139.KL	7155.KL	7161.KL	7172.KL	8176.KL	8206.KL	8869.KL
8966.KL									

Figure 2: Stocks Selected for Portfolio Construction

3.2. Long-run portfolio performance with COVID-19 effect

The estimation window to cluster stocks is three years. Using the three factors discussed in Section 2.2, the optimal clusters using the elbow method is shown in Figure 3. Accordingly, the number of stocks in each cluster with center for each feature is given in Table 1.

The two comprehensively relatively better performing clusters, 3 and 5, are selected for portfolio construction. As shown in Figure 4 for the nine-year performance of the corresponding nineteen-stock portfolio, in nine years, the invested amount of 10,000 ringgits accumulated to 52,652 ringgits, with an annual return of 22.71%, a standard deviation of 27.81%, and a Sharpe ratio of 0.71, based on the risk-free rate of 3% from the Malaysia treasury bills rate 2% and the bond yield 3.50% in January 2010.

In the nineteen-stock portfolio, the stock with ticker 0146.KL is an outlier where the share price increased dramatically in a short period of time due to stock split. After removing the outlier, the performance of the resulting eighteen-stock portfolio is given in Figure 5, which shows that, during the nine years, the invested amount of 10,000 ringgits accumulated to 30,942 ringgits, with an annual returnof 16.50%, a standard deviation of 27.18%, and a Sharpe ratio of 0.50.

Similarly, centers of clusters and the number of stocks when the estimation window is five years is given in Table 2.

Table 2: Centers of Clusters for an Estimation Window of Five Years

Cluster	**AAT**	μ	μ/σ	**Stocks**
0	0.605298	0.061318	0.084933	242
1	1.792953	0.447732	0.802008	134
2	1.458035	13.568424	0.452973	2
3	1.057861	0.302133	0.489820	288
4	6.658005	1.466990	1.181057	5
5	3.435583	1.063792	1.087505	24

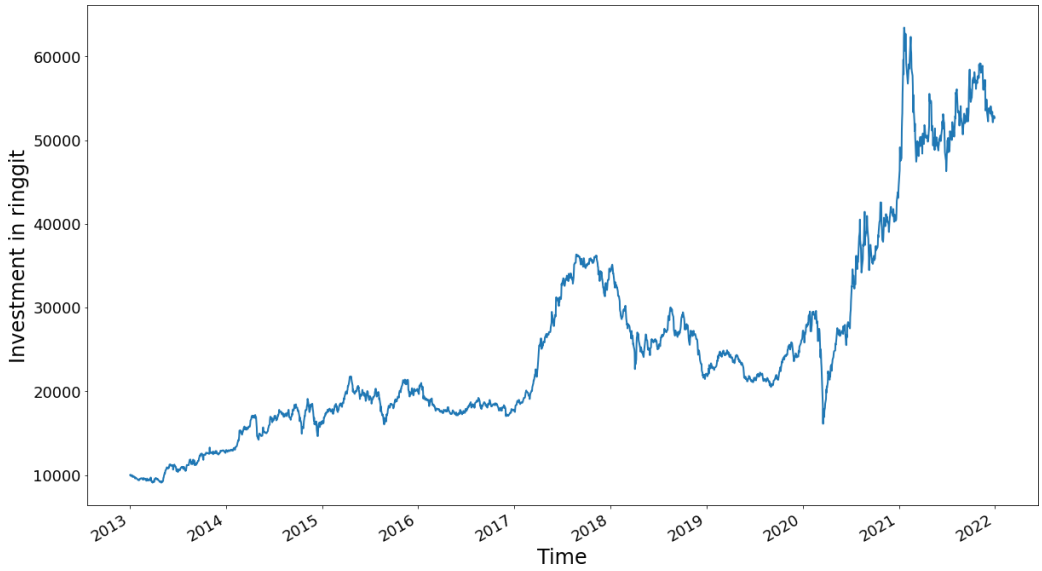

Figure 4: Performance of a Nineteen-Stock Portfolio with an Initial Investment of RM 10,000 in January 2013

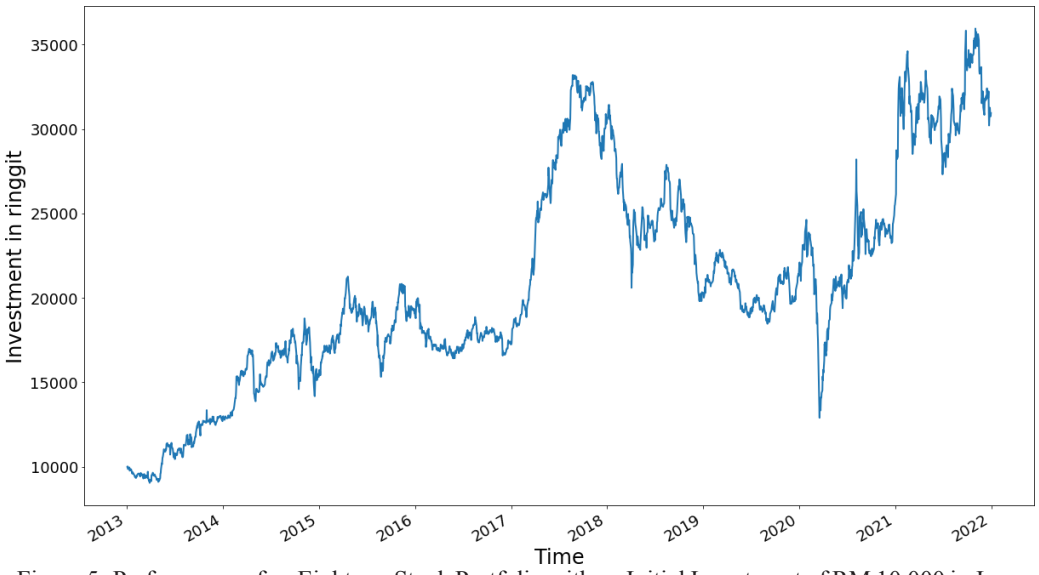

Figure 5: Performance of an Eighteen-Stock Portfolio with an Initial Investment of RM 10,000 in January 2013

In this case, the two relatively better performing clusters, 4 and 5, are selected for portfolio construction. For the seven-year performance of the corresponding twenty-nine-stock portfolio, as shown in Figure 6, in seven years, the invested amount of 10,000 ringgits accumulated to 20,708 ringgits, with

anannual return of 12.35%, a standard deviation of 18.57%, and a Sharpe ratio of 0.50. The sudden declinein 2020 is mainly due to the effect of COVID-19.

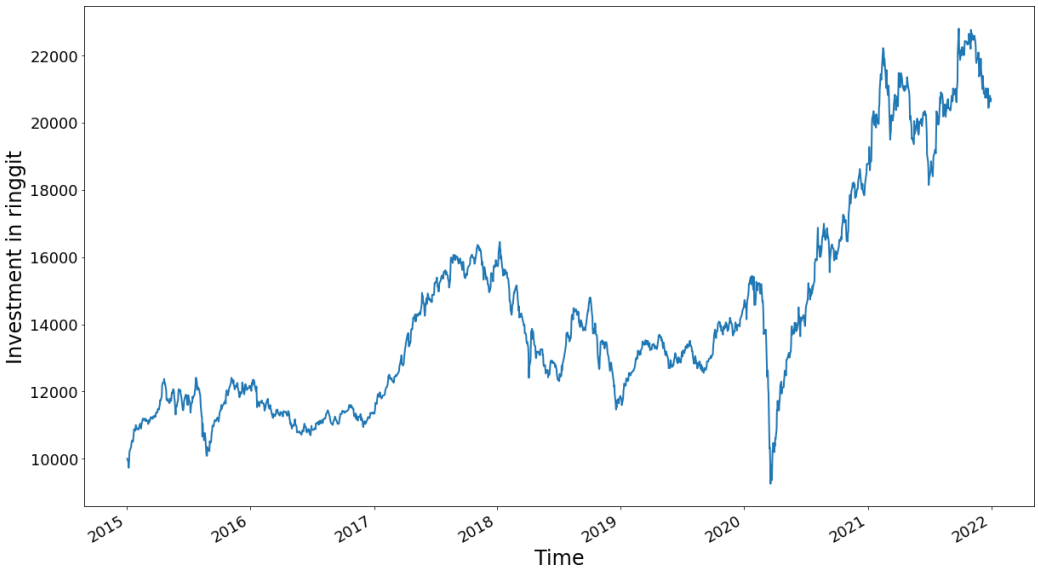

Figure 6: Performance of a Twenty-Nine-Stock Portfolio with an Initial Investment of RM 10,000 in January 2015

The nine-year portfolio performance is based on three-year clustering while the seven-year portfolio performance is based on five-year clustering. In the case of a nine-year trajectory as displayed in Figure 5, the portfolio performance is optimistic with a Sharpe ratio of 0.5. The portfolio is also relatively stable during the COVID-19 shock. In the case of a seven-year trajectory as shown in Figure 6, the portfolio performance is also competitive with the same Sharpe ratio. Meanwhile, the sharp decline during the COVID-19 shock suggests that more clustering factors may be incorporated to sustain the financial shocks.

3.3. Short-Run Portfolio Performance with COVID-19 Effect

In March 2020, the global stock markets crashed due to the first major wave of COVID-19 [Contessi and De Pace, 2021]. The Malaysian stock market is not an exception as shown in Figure 1.

To additionally test the performance of the portfolios constructed before the COVID-19 pandemic according to the proposed strategy based on the *k*-means clustering, a portfolio is constructed on the first day of January 2019 and the performance is then measured from January 2019 to December 2021 as demonstrated in Figure 7. The invested amount of 10,000 ringgits accumulated to 18403 ringgits, with an annual return of 23.89%, a standard deviation of 24.30%, and a Sharpe ratio of 0.87.

Similarly, for a portfolio constructed on the first day of January 2020, as shown in Figure 8, from January 2020 to December 2021, the invested amount of 10,000 ringgits accumulated to 14529 ringgits, with an annual return of 23.06%, a standard deviation of 27.53%, and a Sharpe ratio of 0.74.

Although there is a sharp decline in March 2020, the portfolios perform better in the long run. To sum up, in this paper, the k-means clustering algorithm is employed to identify promising stocks for forming optimized portfolios. The equity portfolios are constructed by using estimation windows of three and five years, with performance being presented in Figures 5 and 6, respectively. To further ascertain the pandemic effect on portfolios formed right before the global stock market crash, portfolios are additionally

constructed in January 2019 and 2020, with performance of the portfolios being shown in Figures 7 and 8, respectively. The portfolios are hence formed at different time periods with different estimation windows for clustering to provide reasonable robustness check. Considering the KLSE index performance from 2013 to 2021, the proposed portfolio construction methodology performs comparatively well with an average Sharpe ratio of 0.7. Consequently, the k-means algorithm of machine learning with the given clustering factors may provide an alternative tool worth considering to identify optimistic stocks for investment.

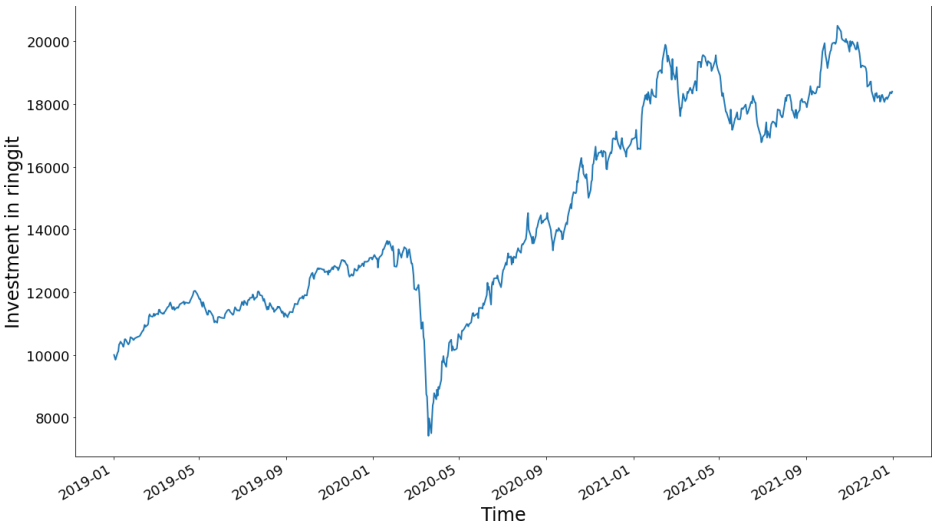

Figure 7: Performance of a Twenty-One-Stock Portfolio from 2019 to 2021

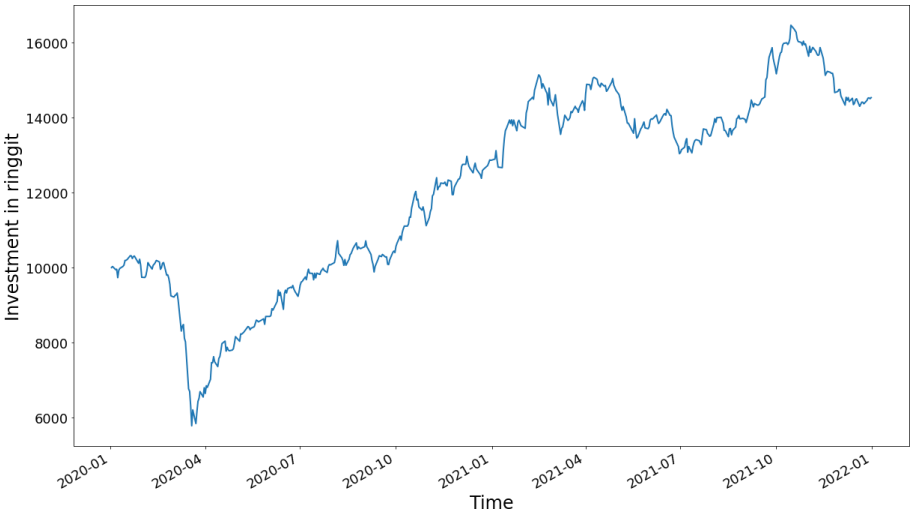

Figure 8: Performance of a Twenty-One-Stock Portfolio from 2020 to 2021

4. Conclusion

Selecting stocks to form a portfolio from a large number of active stocks in the market is a critical investment decision. In this paper, stocks are clustered into different groups with the k-means clustering algorithm. The clusters of stocks with higher risk-adjusted returns are selected for constructing portfolios. There are eighteen and twenty-nine such stocks in case of three and five years of estimation windows, respectively. The results demonstrate high economic gains for an investor with a Sharpe ratio of 0.5 in both cases. The performance is further improved when more data are used to identify optimistic stocks as documented in Section 3.3, where the Sharpe ratios are 0.87 and 0.74 when the portfolios are formed in January 2019 and 2020, respectively, with the exception of the COVID-19 collapse in March 2020 where the constructed portfolios could not sustain the financial shock. In addition, an investor is able to select stocks for portfolio construction from the list of promising stocks only instead of the total 695 stocks. It simplifies and improves the stock selection process. The proposed stock selection strategy may also provide an alternative to be considered for individual or retail investors who normally form concentrated portfolios as they have financial limitations to invest in a large number of stocks.

5. Acknowledgements

This research has been supported by the SDEC (Sarawak Digital Economy Corporation) Translational Research Grant 2022 (Grant Number: A1-CURTIN) and partially by CMHDR (Curtin Malaysia Higher Degree by Research) Grant, which are greatly appreciated.

References

Aslam, B. and Zhang, C. (2022). A strengthened solution to option manipulation. *INFOR: Information Systems and Operational Research*, 60(3):407–427.

Barberis, N. (2013). The psychology of tail events: progress and challenges. *American Economic Review*, 103(3):611–16.

Brealey, R. A., Myers, S. C., Allen, F., and Mohanty, P. (2020). *Principles of corporate finance*. TataMcGraw-Hill Education.

Cao, J., Han, B., and Wang, Q. (2017). Institutional investment constraints and stock prices. *Journal ofFinancial and Quantitative Analysis*, 52(2):465–489.

Contessi, S. and De Pace, P. (2021). The international spread of covid-19 stock market collapses. *Finance Research Letters*, 42:101894.

Guerard, J. B., Markowitza, H., and Xu, G. (2020). Earnings forecasting in a global stock selection model and efficient portfolio construction and management. In *Handbook of Applied Investment Research*, pages 75–85. World Scientific.

Gupta, T. and Kelly, B. (2019). Factor momentum everywhere. *The Journal of Portfolio Management*, 45(3):13–36.

Hull, J. (2021). *Machine Learning in Business: An Introduction to the World of Data Science*. Independently Published.

Klaas, J. (2019). *Machine learning for finance: principles and practice for financial insiders*. Packt Publishing Ltd.

Marvin, K. (2015). Creating diversified portfolios using cluster analysis. *Princeton University*.

Mondal, S. S., Mohanty, S. P., Harlander, B., Koseoglu, M., Rane, L., Romanov, K., Liu, W.-K., Hatwar, P., Salathe, M., and Byrum, J. (2019). Investment ranking challenge: Identifying the best performing stocks based on their semi-annual returns.

Ndikum, P. (2020). Machine learning algorithms for financial asset price forecasting. *arXiv preprint arXiv:2004.01504*.

Ren, Z. (2005). *Portfolio construction using clustering methods*. PhD thesis, Worcester Polytechnic Institute Worcester.

Rozin, P. and Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personalityand social psychology review*, 5(4):296–320.

Soleymani, F. and Vasighi, M. (2022). Efficient portfolio construction by means of CVaR and k-means++ clustering analysis: Evidence from the NYSE. *International Journal of Finance & Economics*, 27(3):3679–3693.

Tan, Z., Yan, Z., and Zhu, G. (2019). Stock selection with random forest: An exploitation of excess returnin the Chinese stock market. *Heliyon*, 5(8):e02310.

Wiest, T. (2022). Momentum: what do we know 30 years after Jegadeesh and Titman's seminal paper? *Financial Markets and Portfolio Management*, pages 1–20.

Wu, D., Wang, X., and Wu, S. (2022). Construction of stock portfolios based on k-means clustering of continuous trend features. *Knowledge-Based Systems*, 252:109358.