“Analyzing Empirical Returns and Risks of Factor Investing Strategies”

Armin Paul Allado and Yutaka Higashiyama

August 2, 2021

Contents

1. Introduction

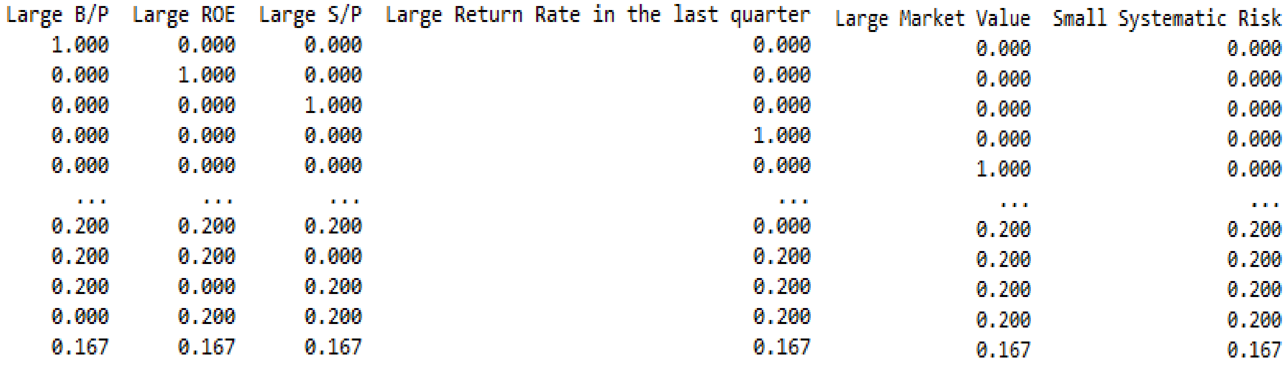
Factor investing is an investment approach that involves allocating portfolio weights to stocks with quantifiable firm characteristics (known as factors). Factor investing can be employed to earn significant risk-adjusted returns in the stock market. For instance, Banz (1981) showed that higher risk-adjusted returns (known as alpha) can be generated from size factor. Small-cap stocks earn on average higher risk-adjusted returns than large-cap stocks. Jegadeesh and Titman (1983) described how momentum factor generates excess returns as current winning stocks tend to outperform current losing stocks in the short-term. Fama and French (1998) illustrated the role of value factor in driving risk-adjusted returns. Value stocks (those with high price-to-book value ratio) tend to outperform growth stocks.

In the past years, portfolio managers and academicians discovered more and more factors that are shown to beat the average stock market returns. Harvey, Liu and Zhu (2016) described more than 300 factors published in top academic journals. Amidst its prevalence, factor investing comes with certain limitations. Arnott et. al. (2019) exposed blunders that adversely affect the profitability of factor investing. For instance, factor investing results are exaggerated due to data mining and unrealistic trading cost expectations. Moreover, factor returns experience substantial downside shocks during market crashes. Factor returns also become more correlated over time, diminishing its diversification effect.

This study aims to extend the factor investing literature by examining the dependencies between well-known factors and common performance metrics on portfolio risk and return. Furthermore, this study will investigate whether a cluster of factor strategies can generate significant risk-adjusted returns over time. Finally, this research will build a portfolio return and risk model using a variety of supervised learning algorithms.

2. Data Source

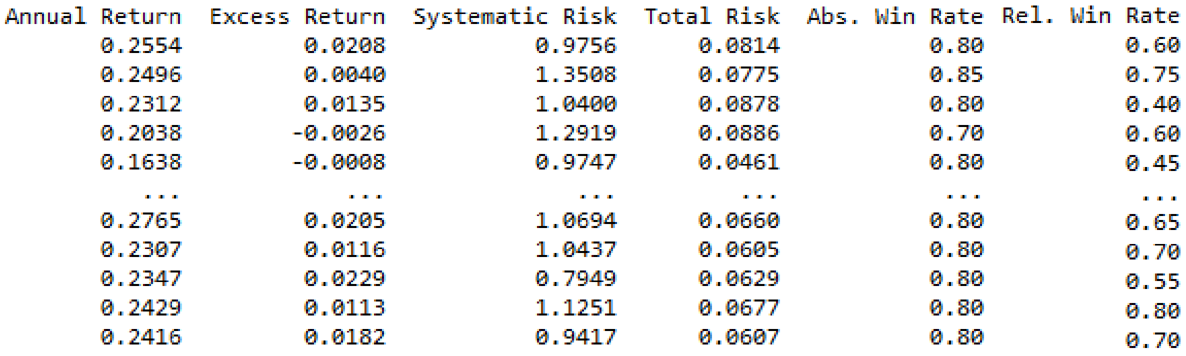
The dataset from Liu and Yeh (2015) will be used to analyze the profitability of factor investing strategies. The dataset contains the resulting returns and risks of different portfolio allocations to five widely known factors (namely *Value*, *Growth*, *Momentum*, *Size* and *Beta*) as shown in Tables 1 and 2:



Large B/P	Large ROE	Large S/P	Large Return Rate in the last quarter	Large Market Value	Small Systematic Risk
1.000	0.000	0.000	0.000	0.000	0.000
0.000	1.000	0.000	0.000	0.000	0.000
0.000	0.000	1.000	0.000	0.000	0.000
0.000	0.000	0.000	1.000	0.000	0.000
0.000	0.000	0.000	0.000	1.000	0.000
...	...	...	...	...	...
0.200	0.200	0.200	0.000	0.200	0.200
0.200	0.200	0.000	0.200	0.200	0.200
0.200	0.000	0.200	0.200	0.200	0.200
0.000	0.200	0.200	0.200	0.200	0.200
0.167	0.167	0.167	0.167	0.167	0.167

Table 1. Different Portfolio Allocations to Five Common Factors

Table 1 provides different portfolio allocations to five different factors. For instance, the first row is the strategy with 100% allocation to Value Factor (as estimated by investing in stocks with high Book-to-Price or B/P ratio). Note that Large B/P and Large S/P estimate the Value factor. Large ROE serves as proxy to Growth factor. Large Return Rate in the last quarter approximates the Momentum factor. Large Market Value and Small Systematic Risk estimate the Size and Beta factors, respectively.



Annual Return	Excess Return	Systematic Risk	Total Risk	Abs. Win Rate	Rel. Win Rate
0.2554	0.0208	0.9756	0.0814	0.80	0.60
0.2496	0.0040	1.3508	0.0775	0.85	0.75
0.2312	0.0135	1.0400	0.0878	0.80	0.40
0.2038	-0.0026	1.2919	0.0886	0.70	0.60
0.1638	-0.0008	0.9747	0.0461	0.80	0.45
...	...	...	...	...	...
0.2765	0.0205	1.0694	0.0660	0.80	0.65
0.2307	0.0116	1.0437	0.0605	0.80	0.70
0.2347	0.0229	0.7949	0.0629	0.80	0.55
0.2429	0.0113	1.1251	0.0677	0.80	0.80
0.2416	0.0182	0.9417	0.0607	0.80	0.70

Table 2. Portfolio Return and Risk of Factor Investing Strategies

Table 2 presents the resulting returns and risks of portfolio allocation weights to five portfolio factors in Table 1. For instance, the strategy of allocating 100% to Value factor in row 1 of Table 1 will yield an annual return of 25.54%, excess return of 2.08%, systematic risk of 0.9756, total risk of 0.0814, absolute win rate of 80%, and relative win rate of 60%.

Book Value-to-Price Ratio (*Large B/P*) and Sales-to-Price Ratio (*Large S/P*) serve as proxy for the *Value Factor*. The value factor stems from the observation that the return rates of undervalued stocks tend to outperform those of overvalued stocks. A higher Book Value-to-Price Ratio or Sales-to-Price Ratio implies a higher probability that a stock is undervalued.

The Large Return on Equity (*Large ROE*) is used as an estimate for the *Growth Factor*. The growth factor pertains to the observation that the return rates of stocks of profitable firms tend to outperform those of non-profitable firms. ROE is often used to assess a firm's profitability. A higher ROE implies that the firm is more profitable.

Momentum Factor depicts the tendency of stocks to follow previously observed trend. For instance, if the recent return of the stock is higher than normal, it will continue to trend higher in the short-term. As the rebalance period in computing portfolio returns in the data is quarterly, the *Return Rate in the last quarter* is used as a proxy to measure degree of momentum.

Market Capitalization estimates the *Size Factor*, which pertains to the observation that stocks with small market capitalization tend to outperform those with large market capitalization. Moreover, the smaller the size of the firm, the less traded the stock tends to be (i.e., the stock is less liquid) and the higher the risk of holding the stock.

Risk factor *Beta (β)* assesses the small systematic risk of the stock, particularly how the stock returns fluctuate with market benchmark returns. If β is greater than 1, the stock's fluctuation is greater than the benchmark and has a higher volatility risk, and vice versa.

Six performance indicators (*Annual Return*, *Excess Return*, *Systematic Risk*, *Total Risk*, *Absolute Win Rate*, and *Relative Win Rate*) are considered as labels. *Annual Return* is calculated from the following formula:

$$\mathrm{ARR} = \left((1 + R)^{\frac{1}{t}}\right) - 1 \tag{1}$$

where *R* is the accumulated return rates and *t* is the duration in years.

Excess return is estimated from the regression coefficient value *α* in the following regression equation:

$$R_{\mathrm{i}} - R_{\mathrm{f}} = \alpha_i + \beta_i(R_{\mathrm{m}} - R_{\mathrm{f}}) \tag{2}$$

If the excess return *α* > 0, it shows that portfolio *i* is generating returns above the general return of the stock market.

Systematic Risk is derived from *β* in Eq. (2). A higher value of *β* implies a higher systematic risk of the portfolio.

Total Risk σ is measured by the standard deviation of portfolio return. It refers to the volatility of portfolio return rate in a given period of time. The large volatility of portfolio return rate is captured by a high standard deviation of portfolio return.

Absolute Win Rate (Win_{abs}*)* is determined by the following formula:

$$\mathrm{Win}_{\mathrm{abs}} = \frac{n_1}{N} \tag{3}$$

where n_1 is the number of portfolio holding periods when the portfolio return rate is above 0 and *N* is the total number of portfolio holding periods.

Relative Win Rate (Win_{rel}*)* is determined by the following formula:

$$\mathrm{Win}_{\mathrm{rel}} = \frac{n_2}{N} \tag{4}$$

where n_2 is the number of portfolio holding periods when the portfolio return rate is above the general stock market return.

Finally, the dataset is divided into four time periods:

Period	Date
Period 1	1990/9/30 to 1995/6/30
Period 2	1995/9/30 to 2000/6/30
Period 3	2000/9/30 to 2005/6/30
Period 4	2005/9/30 to 2010/6/30
All Periods	1990/9/30 to 2010/6/30

Table 3. Historical Time Period

The entire dataset covers the twenty-year period from 30 September 1990 to 30 June 2010. Each of the four subperiods shows the portfolio returns and risk of different factor allocations when implemented in the given period. The data also assumes that the portfolio is rebalanced every quarter (3 months).

3. Methodology

The performance indicators on returns and risks will be scaled to account for differences in units. *Mutual Information (MI)* between the six proxy estimates for the factor strategies (features) and each of the performance indicators (labels) will be computed to determine the degree of dependencies between the features and the labels. A high Mutual Information signals that a given feature contains a good amount of information on a performance indicator and provides insight on the key factors that drive the movement of the performance indicator.

K-Means Clustering will be used to find natural clustering of different factor allocations based on their performance indicators for each of the subperiods. Finally, a variety of supervised learning methods will be implemented to build a portfolio return and risk model. The model will determine the fair value of each of the performance indicators given different factor allocations.10-Fold Cross-Validated R^2 will be used to evaluate and select the best model.

Methodology flowchart is presented in Figure 1, and the set-ups of these methods will be discussed in more depth in subsequent parts:

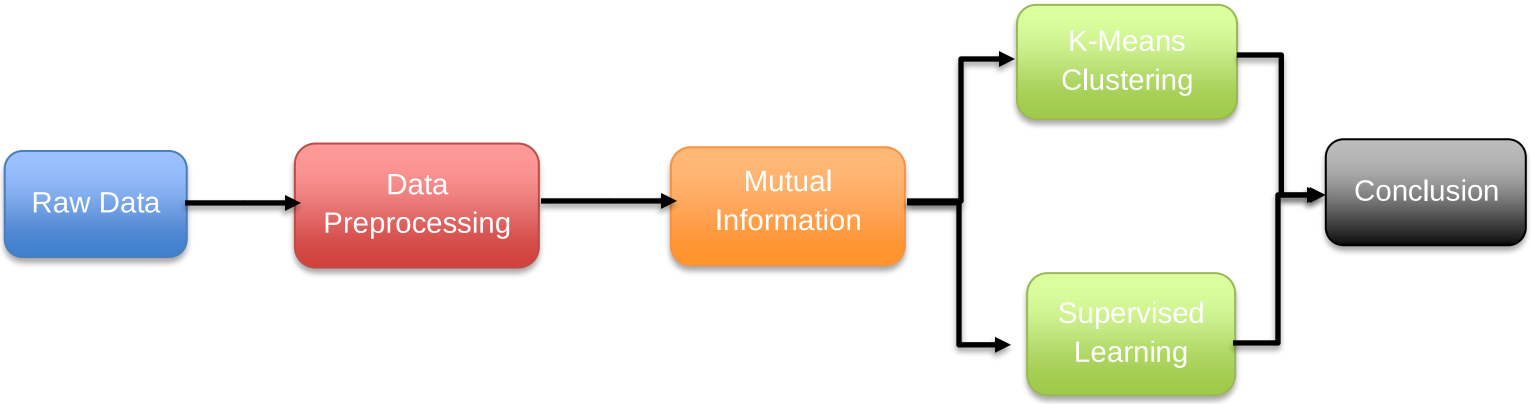


Figure 1. Methodology Flowchart

4. Data Preprocessing

Each of the performance indicators was scaled and normalized to account for differences in units. Failure to scale the labels can distort the resulting portfolio return and risk model using supervised learning algorithms. Note that the six features in the raw data are already expressed in percentage form and the corresponding features for each data point sum up to 1. Hence, no data preprocessing is necessary for the features.

5. Mutual Information

Mutual information between the six features and each of the performance indicators was determined to gain insight on significant factors that drive the movement of the performance indicators. Mutual information result and analysis are presented in Tables 4 to 9:

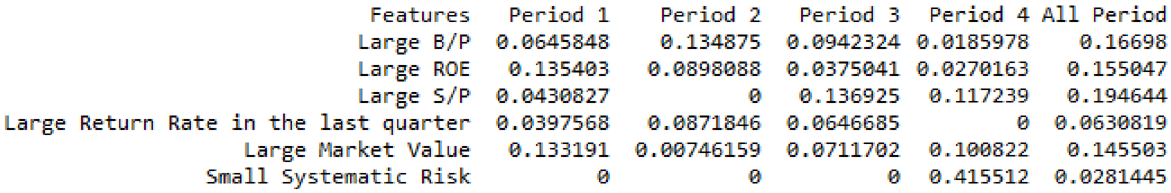
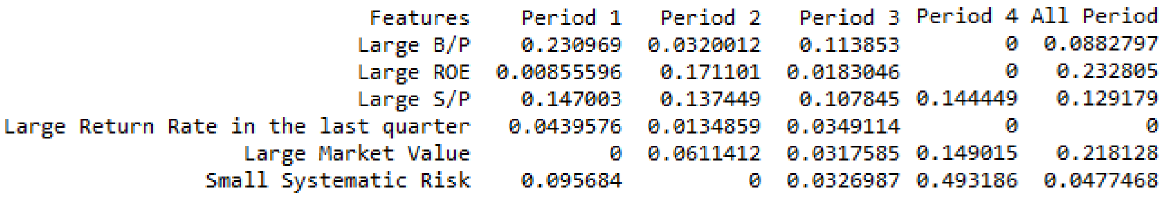

```
                          Features   Period 1    Period 2   Period 3  Period 4 All Period
                         Large B/P 0.0645848    0.134875 0.0942324 0.0185978    0.16698
                         Large ROE  0.135403   0.0898088 0.0375041 0.0270163   0.155047
                         Large S/P 0.0430827           0  0.136925  0.117239   0.194644
Large Return Rate in the last quarter 0.0397568   0.0871846 0.0646685         0  0.0630819
                Large Market Value  0.133191  0.00746159 0.0711702  0.100822   0.145503
              Small Systematic Risk         0           0         0  0.415512  0.0281445
```

Table 4. Mutual Information between Six Factor Strategies and Annual Returns

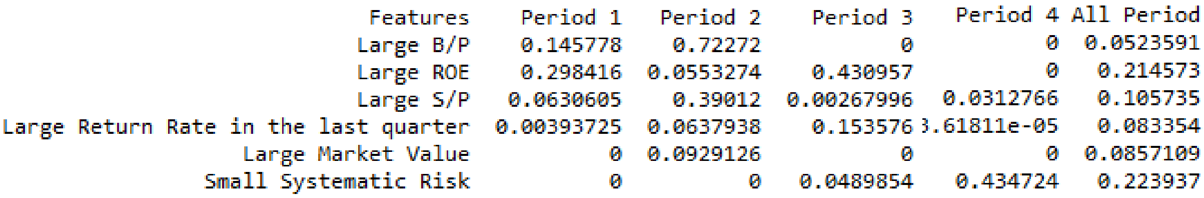
Large B/P, *Large ROE* and *Large Market Value* proved to have consistently positive and high mutual dependence with annual returns than other features across all subperiods. *Small systematic risk* only showed to have mutual dependence with annual returns in recessionary Period 4.



```
                             Features   Period 1  Period 2  Period 3 Period 4 All Period
                            Large B/P   0.230969 0.0320012  0.113853        0  0.0882797
                            Large ROE 0.00855596  0.171101 0.0183046        0   0.232805
                            Large S/P   0.147003  0.137449  0.107845 0.144449   0.129179
Large Return Rate in the last quarter  0.0439576 0.0134859 0.0349114        0          0
                   Large Market Value          0 0.0611412 0.0317585 0.149015   0.218128
                Small Systematic Risk   0.095684         0 0.0326987 0.493186  0.0477468
```

Table 5. Mutual Information between Six Factor Strategies and Excess Returns

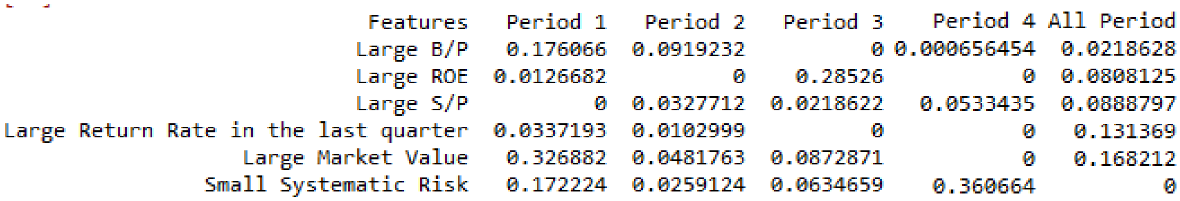
Large S/P had sustainably high mutual dependence with excess returns. Moreover, only *Large S/P, Large Market Value*, and *Small Systematic Risk* contained information about excess returns in the 4th subperiod. *Large Return Rate in the last quarter* had consistently low mutual information with excess returns.



```
                             Features   Period 1  Period 2   Period 3   Period 4 All Period
                            Large B/P   0.145778   0.72272          0          0  0.0523591
                            Large ROE   0.298416 0.0553274   0.430957          0   0.214573
                            Large S/P  0.0630605   0.39012 0.00267996  0.0312766   0.105735
Large Return Rate in the last quarter 0.00393725 0.0637938   0.1535763.61811e-05   0.083354
                   Large Market Value          0 0.0929126          0          0  0.0857109
                Small Systematic Risk          0         0  0.0489854   0.434724   0.223937
```

Table 6. Mutual Information between Six Factor Strategies and Systematic Risk

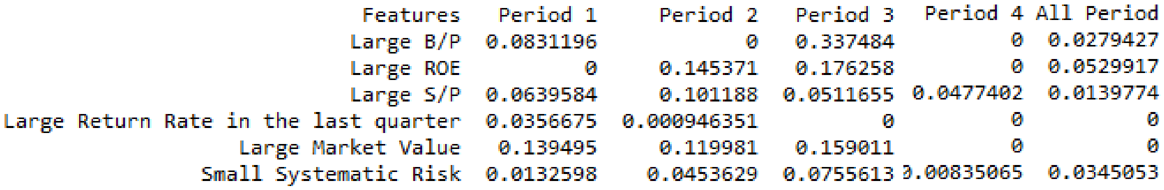
Large S/P proved to have consistently positive mutual dependence with systematic risk across all subperiods. *Large Return Rate in the last quarter* also showed some mutual dependence with systematic risk although at a lesser degree than *Large S/*P. Mutual information of other features varied greatly across each period.



Features	Period 1	Period 2	Period 3	Period 4	All Period
Large B/P	0.176066	0.0919232	0	0.000656454	0.0218628
Large ROE	0.0126682	0	0.28526	0	0.0808125
Large S/P	0	0.0327712	0.0218622	0.0533435	0.0888797
Large Return Rate in the last quarter	0.0337193	0.0102999	0	0	0.131369
Large Market Value	0.326882	0.0481763	0.0872871	0	0.168212
Small Systematic Risk	0.172224	0.0259124	0.0634659	0.360664	0

Table 7. Mutual Information between Six Factor Strategies and Total Risk

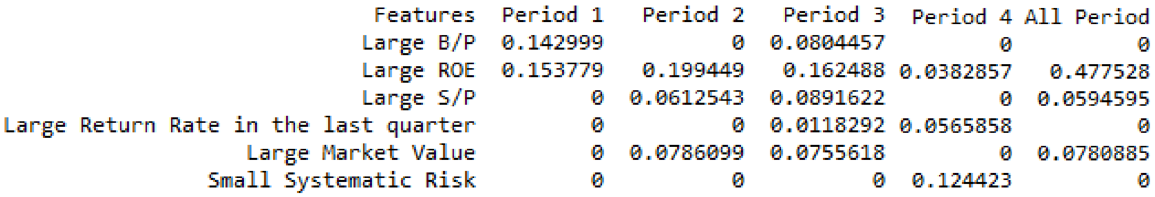
Small systematic risk had consistently high mutual information with total risk, suggesting the importance of measuring the β of each stock in understanding the drivers of the stock's total risk. Other features had inconsistent mutual information with total risk across the four subperiods. Furthermore, only small systematic risk showed to have significant dependence with total risk during recessionary period 4.



Features	Period 1	Period 2	Period 3	Period 4	All Period
Large B/P	0.0831196	0	0.337484	0	0.0279427
Large ROE	0	0.145371	0.176258	0	0.0529917
Large S/P	0.0639584	0.101188	0.0511655	0.0477402	0.0139774
Large Return Rate in the last quarter	0.0356675	0.000946351	0	0	0
Large Market Value	0.139495	0.119981	0.159011	0	0
Small Systematic Risk	0.0132598	0.0453629	0.0755613	0.00835065	0.0345053

Table 8. Mutual Information between Six Factor Strategies and Absolute Win Rate

Large S/P proved to have consistent positive mutual dependence with absolute win rate across the four subperiods. *Small systematic risk* also showed to have sustained mutual dependence with absolute win rate albeit at a lower degree than *Large S/P*.



Features	Period 1	Period 2	Period 3	Period 4	All Period
Large B/P	0.142999	0	0.0804457	0	0
Large ROE	0.153779	0.199449	0.162488	0.0382857	0.477528
Large S/P	0	0.0612543	0.0891622	0	0.0594595
Large Return Rate in the last quarter	0	0	0.0118292	0.0565858	0
Large Market Value	0	0.0786099	0.0755618	0	0.0780885
Small Systematic Risk	0	0	0	0.124423	0

Table 9. Mutual Information between Six Factor Strategies and Relative Win Rate

Large ROE showed to have high and consistent mutual dependence with relative win rate across all subperiods. Other features had volatile and inconsistent mutual information with relative win rate. Only *Large ROE*, *Large Return Rate in the last* *quarter*, and *Small Systematic Risk* proved to have mutual dependence with relative win rate during the contractionary 4th period.

Initial analysis using mutual information revealed the most important feature for each performance indicator:

Performance Indicator	Key Feature
Annual Returns	Large B/P, Large ROE, Large Market Value
Excess Returns	Large S/P
Systematic Risk	Large S/P
Total Risk	Small Systematic Risk
Absolute Win Rate	Large S/P
Relative Win Rate	Large ROE

Table 10. Key Feature for the Six Performance Indicators

6. **Clustering of Factor Allocations**

K-Means Clustering was implemented with the scaled six performance indicators served as components. The number of clusters *K* was selected using an elbow diagram. Based on the components for each subperiod, *K* clusters were formed. The average Sharpe Ratios (*Excess Return* divided by *Systematic Risk*) for each cluster were computed. Note that Sharpe Ratio was calculated to determine which of the formed clusters generate the highest excess return per unit of risky investment taken. The cluster with the highest Sharpe Ratio will be selected in the next subperiod. Moreover, its corresponding Sharpe Ratio in the next subperiod will be compared with the previous period Sharpe Ratio.

The goal of K-Means Clustering implementation is to test whether the same cluster of factors that delivered the highest Sharpe Ratio in one subperiod will also generate a high Sharpe Ratio in the next subperiod. This means that the same strategy of weight allocation (for instance, 50% in *Large ROE* stocks, 25% in *Large B/P* stocks, and 25% in *Large Market Value* stocks) that worked during the previous 5 years will continue to work in the next 5 years.

K-Means Clustering in Period 1

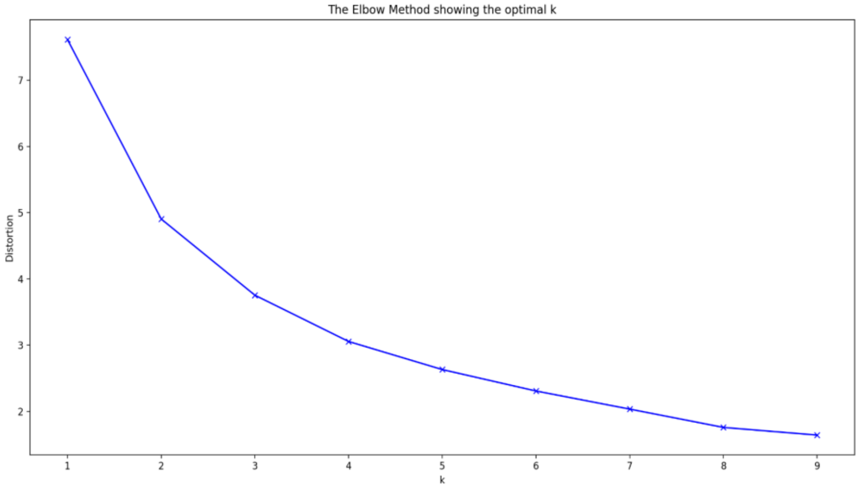
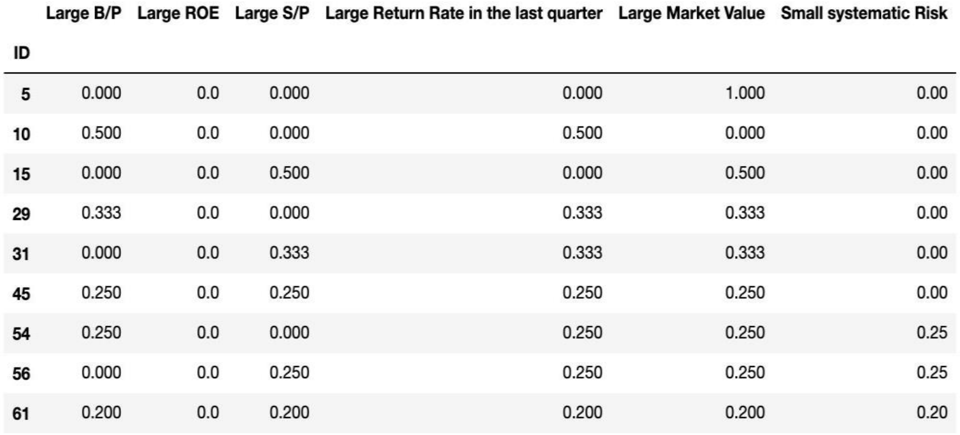


Figure 2. Elbow Method Result for K-Means Clustering in Period 1

From the elbow diagram that shows the Total Within-Cluster Sum-of-Squares for a given number of cluster *k*, the tipping point occurred at the point when *k* = 4, and 4 clusters were formed in the analysis.



ID	Large B/P	Large ROE	Large S/P	Large Return Rate in the last quarter	Large Market Value	Small systematic Risk
5	0.000	0.0	0.000	0.000	1.000	0.00
10	0.500	0.0	0.000	0.500	0.000	0.00
15	0.000	0.0	0.500	0.000	0.500	0.00
29	0.333	0.0	0.000	0.333	0.333	0.00
31	0.000	0.0	0.333	0.333	0.333	0.00
45	0.250	0.0	0.250	0.250	0.250	0.00
54	0.250	0.0	0.000	0.250	0.250	0.25
56	0.000	0.0	0.250	0.250	0.250	0.25
61	0.200	0.0	0.200	0.200	0.200	0.20

Table 11. Cluster of Factor Weights with the Highest Sharpe Ratio in Period 1

Table 11 shows the elements of the cluster that generated the highest Sharpe Ratio in Period 1. The factor weight allocations (indexed by their ID numbers) in the selected cluster are presented. For instance, index 5 pertains to the factor weight allocation of 100% in *Large Market Value*. A close look at the selected cluster suggests that some exposure to Size factor (proxied by *Large Market Value*) and

Momentum factor (proxied by *Large Return Rate in the last quarter*) yielded the highest Sharpe Ratio in Period 1.

Selected Factor Index	Sharpe Ratio
Selected Cluster (Indices 5, 10, 29, 31, 45, 54, 56, 61)	1.662
All 64 Indices	1.278
Difference between Selected Cluster and Average Sharpe Ratio	0.384

Table 12. Comparison of Sharpe Ratios

Table 12 presents the resulting Sharpe Ratio when the formed cluster in Period 1 is selected in subsequent Period 2. The chosen cluster in Period 1 will yield a Sharpe Ratio of 1.662 in Period 2, higher than the average Sharpe Ratio of 1.278 of all the 64 factor-based allocations in Period 2.

K-Means Clustering in Period 2

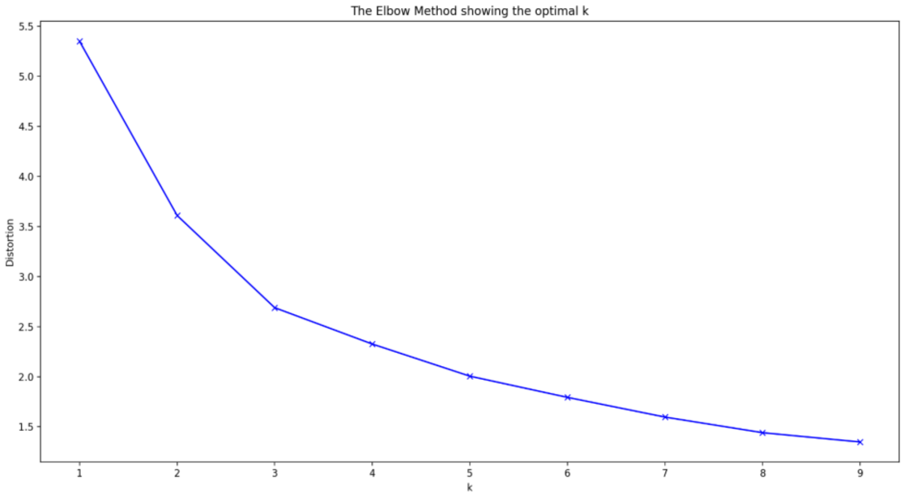
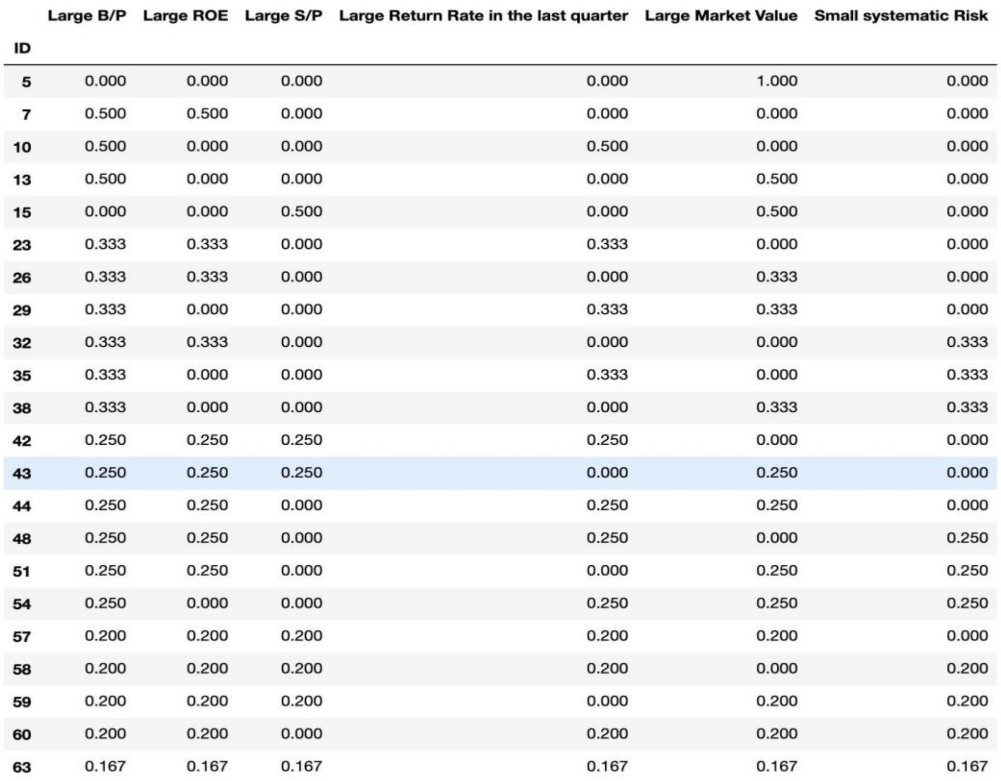


Figure 3. Elbow Method Result for K-Means Clustering in Period 2

Similar with the K-Means Clustering in Period 1, the tipping point occurred at the point when k = 4, and 4 clusters were formed in the analysis.



ID	Large B/P	Large ROE	Large S/P	Large Return Rate in the last quarter	Large Market Value	Small systematic Risk
5	0.000	0.000	0.000	0.000	1.000	0.000
7	0.500	0.500	0.000	0.000	0.000	0.000
10	0.500	0.000	0.000	0.500	0.000	0.000
13	0.500	0.000	0.000	0.000	0.500	0.000
15	0.000	0.000	0.500	0.000	0.500	0.000
23	0.333	0.333	0.000	0.333	0.000	0.000
26	0.333	0.333	0.000	0.000	0.333	0.000
29	0.333	0.000	0.000	0.333	0.333	0.000
32	0.333	0.333	0.000	0.000	0.000	0.333
35	0.333	0.000	0.000	0.333	0.000	0.333
38	0.333	0.000	0.000	0.000	0.333	0.333
42	0.250	0.250	0.250	0.250	0.000	0.000
43	0.250	0.250	0.250	0.000	0.250	0.000
44	0.250	0.250	0.000	0.250	0.250	0.000
48	0.250	0.250	0.000	0.250	0.000	0.250
51	0.250	0.250	0.000	0.000	0.250	0.250
54	0.250	0.000	0.000	0.250	0.250	0.250
57	0.200	0.200	0.200	0.200	0.200	0.000
58	0.200	0.200	0.200	0.200	0.000	0.200
59	0.200	0.200	0.200	0.000	0.200	0.200
60	0.200	0.200	0.000	0.200	0.200	0.200
63	0.167	0.167	0.167	0.167	0.167	0.167

Table 13. Cluster of Factor Weights with the Highest Sharpe Ratio in Period 2

Table 13 lists the elements of the cluster that generated the highest Sharpe Ratio in Period 2. A thorough look at the cluster elements showed that employing a more diversified investment approach across each of the factor-based strategies will yield a higher Sharpe Ratio in Period 2. This observation is contrary to the optimal strategy in Period 1, which calls for a more concentrated exposure to *Size* and *Momentum* factors.

Selected Factor Index	Sharpe Ratio
Selected Cluster (Indices 5, 7,10, 13, 15, 23, 26, 29, 32, 35, 38, 42, 43, 44, 48, 51, 54, 57, 58, 59, 60, 63)	1.920
All 64 Indices	1.472
Difference between Selected Cluster and Average Sharpe Ratio	0.448

Table 14. Comparison of Sharpe Ratios

Table 14 calculates the resulting Sharpe Ratio when the formed cluster in Period 2 is selected in subsequent Period 3. The chosen cluster in Period 2 will yield a Sharpe Ratio of 1.920 in Period 3, higher than the average Sharpe Ratio of 1.472 of all the 64 factor-based allocations in Period 3.

K-Means Clustering in Period 3

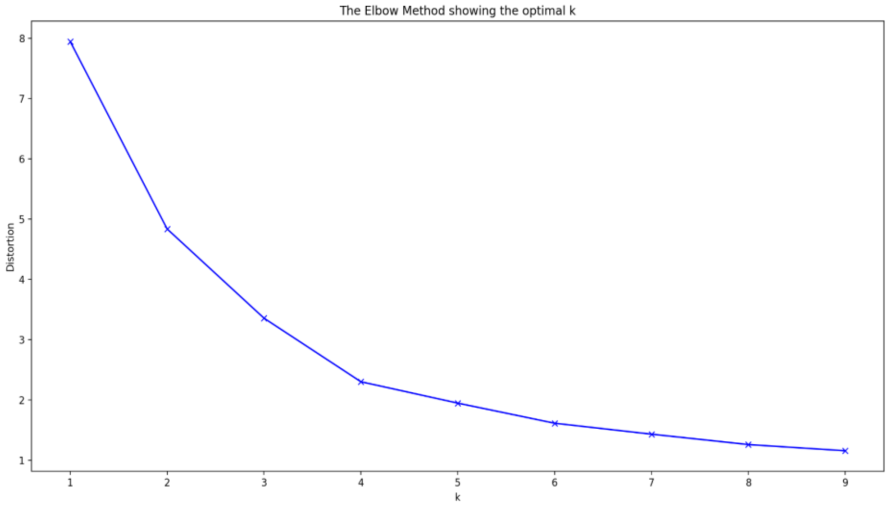
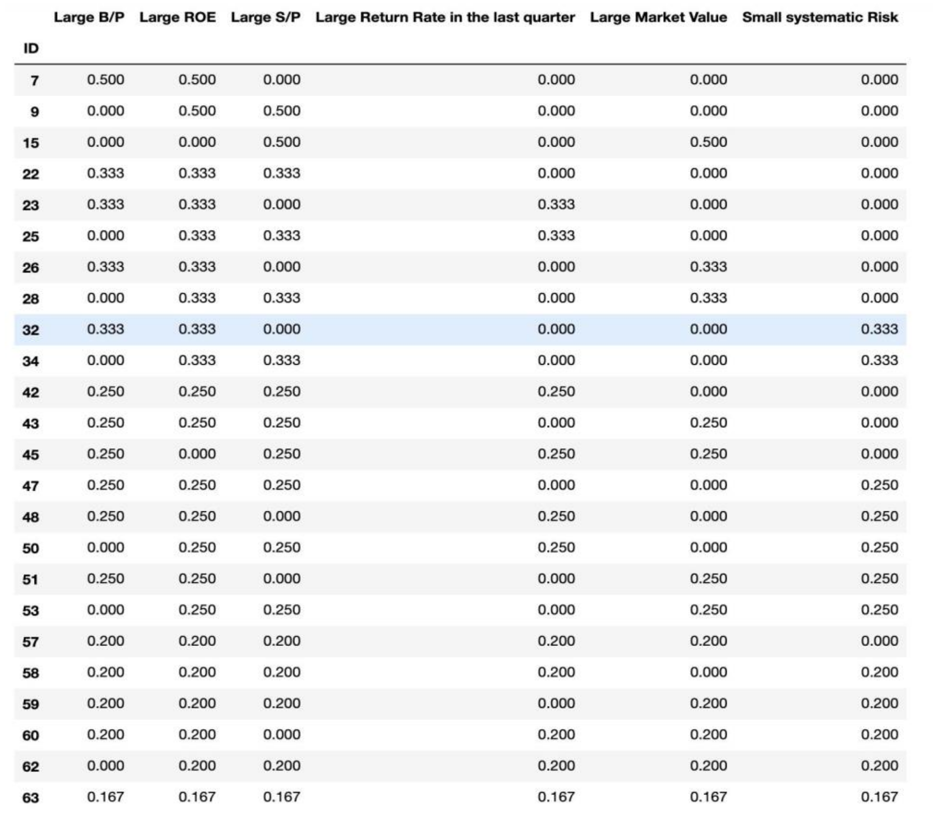


Figure 4. Elbow Method Result for K-Means Clustering in Period 3

Similar with the K-Means Clustering in Periods 1 and 2, the tipping point occurred at the point when *k* = 4, and 4 clusters were formed in the analysis.



ID	Large B/P	Large ROE	Large S/P	Large Return Rate in the last quarter	Large Market Value	Small systematic Risk
7	0.500	0.500	0.000	0.000	0.000	0.000
9	0.000	0.500	0.500	0.000	0.000	0.000
15	0.000	0.000	0.500	0.000	0.500	0.000
22	0.333	0.333	0.333	0.000	0.000	0.000
23	0.333	0.333	0.000	0.333	0.000	0.000
25	0.000	0.333	0.333	0.333	0.000	0.000
26	0.333	0.333	0.000	0.000	0.333	0.000
28	0.000	0.333	0.333	0.000	0.333	0.000
32	0.333	0.333	0.000	0.000	0.000	0.333
34	0.000	0.333	0.333	0.000	0.000	0.333
42	0.250	0.250	0.250	0.250	0.000	0.000
43	0.250	0.250	0.250	0.000	0.250	0.000
45	0.250	0.000	0.250	0.250	0.250	0.000
47	0.250	0.250	0.250	0.000	0.000	0.250
48	0.250	0.250	0.000	0.250	0.000	0.250
50	0.000	0.250	0.250	0.250	0.000	0.250
51	0.250	0.250	0.000	0.000	0.250	0.250
53	0.000	0.250	0.250	0.000	0.250	0.250
57	0.200	0.200	0.200	0.200	0.200	0.000
58	0.200	0.200	0.200	0.200	0.000	0.200
59	0.200	0.200	0.200	0.000	0.200	0.200
60	0.200	0.200	0.000	0.200	0.200	0.200
62	0.000	0.200	0.200	0.200	0.200	0.200
63	0.167	0.167	0.167	0.167	0.167	0.167

Table 15. Cluster of Factor Weights with the Highest Sharpe Ratio in Period 3

Table 15 lists the elements of the cluster that generated the highest Sharpe Ratio in Period 3. Similar with the finding in Period 2, adopting a more diversified investment approach across each of the factor-based strategies will yield a higher Sharpe Ratio in Period 3.

Selected Factor Index	Sharpe Ratio
Selected Cluster (Indices 5, 7, 9, 15, 22, 23, 25, 26, 28, 32, 34, 42, 43, 45, 47, 48, 50, 51, 53, 57, 58, 59, 60, 62, 63)	1.477
All 64 Indices	1.360
Difference between Selected Cluster and Average Sharpe Ratio	0.117

Table 16. Comparison of Sharpe Ratios

Table 16 computes for the resulting Sharpe Ratio when the formed cluster in Period 3 is selected in subsequent Period 4. The chosen cluster in Period 3 will yield a Sharpe Ratio of 1.477 in Period 4, higher than the average Sharpe Ratio of 1.360 of all the 64 factor-based allocations in Period 3.

The implementation of K-Means Clustering across the three subperiods provided some evidence that forming a cluster of factor-based strategies with highest Sharpe Ratio in the current period can generate a higher-than-average Sharpe Ratio in the next 5-year period. However, the difference in Sharpe Ratios declined in Period 3, suggesting that the strategy may experience diminishing returns in the future.

An analysis of cluster elements showed that a more concentrated exposure towards Momentum and Size factor is optimal in Period. In contrast, a more diversified allocation across each of the factor-based strategies works best in realizing a higher Sharpe Ratio in the subsequent periods. Indeed, optimal investment strategies change over time as popular investment strategies may get crowded and have reduced profitability. Finally, the cluster with the highest Sharpe Ratio in periods 2 and 3 contained about 1/3 of total data points. This observation implies that the risk and return profile of factor-based strategies become more similar over the years.

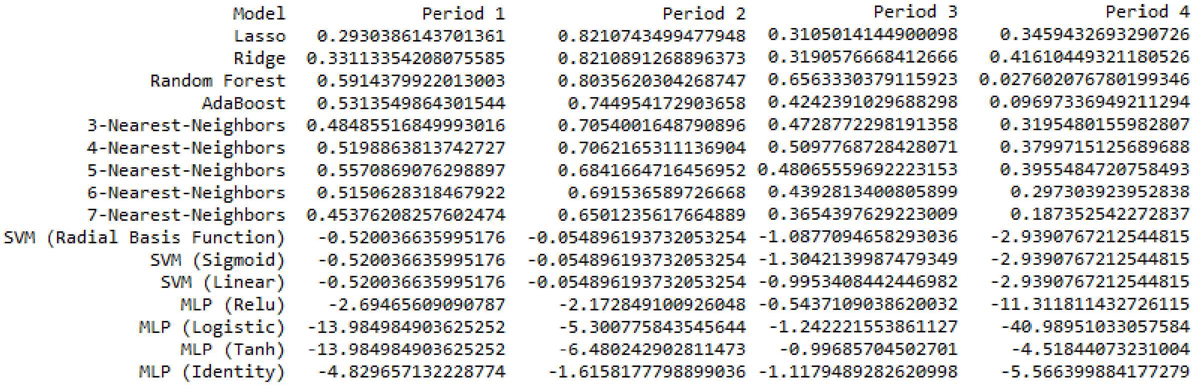
7. Portfolio Return and Risk Model

Supervised learning algorithms were employed to build a reliable portfolio return and risk model. The purpose of the model is to provide a fair estimate of long-term investment return and risk (labels) given different portfolio weight allocations (features) to *Large B/P*, *Large ROE*, *Large S/P*, *Large Return Rate in the last quarter*, *Large Market Value,* and *Small Systematic Riks.*

A variety of supervised learning algorithms were used to train and test the dataset for each of the four subperiods. *Linear regression with feature selection component* (*Lasso* and *Ridge*), *Ensemble Methods* (*Random Forest* and *Adaboost*), *K-Nearest-Neighbors*, *Support Vector Machine (Radial Basis Function, Linear,* and *Sigmoid),* and *Multi-Layer Perceptron (Relu, Logistic, Tanh,* and *Identity)* were implemented to build a regression-based model to reliably estimate the portfolio returns and risks given weight allocations to each of the 6 factor-based strategies. *10-Fold Cross-Validated R^2* was chosen as the performance metric in selecting the supervised learning algorithm to predict future data points.

The final output of the analysis is to determine the best supervised learning algorithm in estimating annual returns, excess returns, systematic risk, and total risk. The selected supervised learning algorithm will be used to estimate the long-term return and risk of different factor weight allocations.

Building a Portfolio Return Model on Annual Returns



Model	Period 1	Period 2	Period 3	Period 4
Lasso	0.2930386143701361	0.8210743499477948	0.3105014144900098	0.3459432693290726
Ridge	0.33113354208075585	0.8210891268896373	0.3190576668412666	0.41610449321180526
Random Forest	0.5914379922013003	0.8035620304268747	0.6563330379115923	0.027602076780199346
AdaBoost	0.5313549864301544	0.744954172903658	0.4242391029688298	0.09697336949211294
3-Nearest-Neighbors	0.48485516849993016	0.7054001648790896	0.4728772298191358	0.3195480155982807
4-Nearest-Neighbors	0.5198863813742727	0.7062165311136904	0.5097768728428071	0.3799715125689688
5-Nearest-Neighbors	0.5570869076298897	0.6841664716456952	0.48065559692223153	0.3955484720758493
6-Nearest-Neighbors	0.5150628318467922	0.691536589726668	0.4392813400805899	0.297303923952838
7-Nearest-Neighbors	0.45376208257602474	0.6501235617664889	0.3654397629223009	0.187352542272837
SVM (Radial Basis Function)	-0.520036635995176	-0.054896193732053254	-1.0877094658293036	-2.9390767212544815
SVM (Sigmoid)	-0.520036635995176	-0.054896193732053254	-1.3042139987479349	-2.9390767212544815
SVM (Linear)	-0.520036635995176	-0.054896193732053254	-0.9953408442446982	-2.9390767212544815
MLP (Relu)	-2.69465609090787	-2.172849100926048	-0.5437109038620032	-11.311811432726115
MLP (Logistic)	-13.984984903625252	-5.300775843545644	-1.242221553861127	-40.98951033057584
MLP (Tanh)	-13.984984903625252	-6.480242902811473	-0.99685704502701	-4.51844073231004
MLP (Identity)	-4.829657132228774	-1.6158177798899036	-1.1179489282620998	-5.566399884177279

Table 16. 10-Fold Cross-Validated R^2 of Supervised Learning Algorithms

Ensemble methods (Random Forest and Adaboost) performed best in predicting annual returns based on different factor weight allocations. However, their cross-validated R^2 scores are diminished during recessionary period 4. This calls for the need to retrain the ensemble algorithms with changes in economic regimes. Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) models fared poorly across all subperiods.

Large B/P	Large ROE	Large S/P	Large Return Rate in last quarter	Large Market Value	Small Systematic Risk	Predicted 20-year Annual Return
100%	0%	0%	0%	0%	0%	**14.061%**
0%	100%	0%	0%	0%	0%	**15.301%**
0%	0%	100%	0%	0%	0%	**16.601%**
0%	0%	0%	100%	0%	0%	**10.616%**
0%	0%	0%	0%	100%	0%	**10.006%**
0%	0%	0%	0%	0%	100%	**8.607%**
16.67%	16.67%	16.67%	16.67%	16.67%	16.67%	**17.094%**

Table 17. Random Forest Estimate of 20-Year Annual Returns with various Factor Allocations

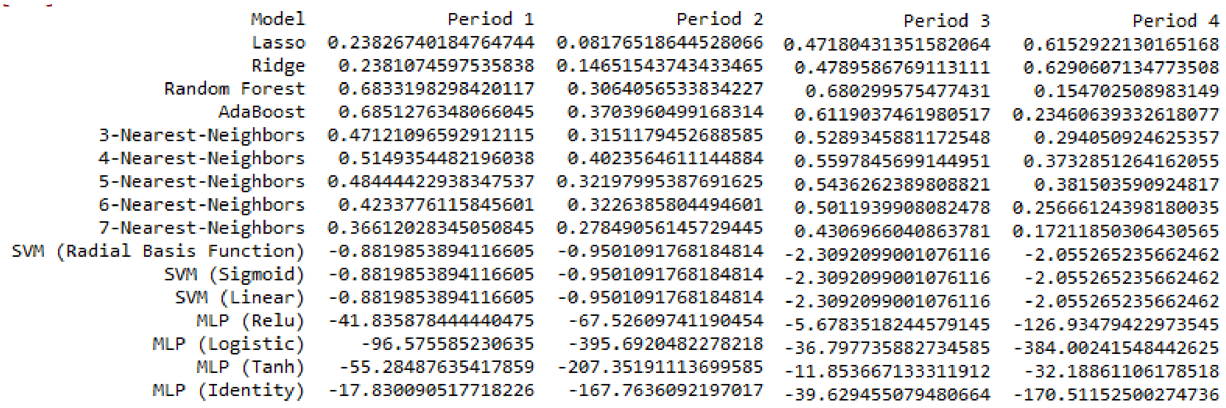
Random Forest algorithm estimated an equally weighted portfolio across the six factor-based strategies to generate the highest 20-year annual return among the 7 scenarios. It predicted a portfolio fully allocated to *Small Systematic Risk* strategy to yield the lowest 20-year annual return.

Large B/P	Large ROE	Large S/P	Large Return Rate in last quarter	Large Market Value	Small Systematic Risk	Predicted 20-year Annual Return
100%	0%	0%	0%	0%	0%	**14.038%**
0%	100%	0%	0%	0%	0%	**15.318%**
0%	0%	100%	0%	0%	0%	**16.272%**
0%	0%	0%	100%	0%	0%	**9.453%**
0%	0%	0%	0%	100%	0%	**9.345%**
0%	0%	0%	0%	0%	100%	**8.725%**
16.67%	16.67%	16.67%	16.67%	16.67%	16.67%	**16.593%**

Table 18. Adaboost Estimate of 20-Year Annual Returns with various Factor Allocations

Similar with Random Forest algorithm, Adaboost forecasted an equally weighted portfolio across the six factor-based strategies to generate the highest 20-year annual return among the 7 scenarios. It estimated a portfolio 100% allocated to *Small Systematic Risk* strategy to yield the lowest 20-year annual return.

Building a Portfolio Return Model on Excess Returns



```
                      Model             Period 1             Period 2             Period 3             Period 4
                      Lasso  0.23826740184764744  0.08176518644528066  0.47180431351582064   0.6152922130165168
                      Ridge   0.2381074597535838  0.14651543743433465   0.4789586769113111   0.6290607134773508
              Random Forest   0.6833198298420117   0.3064056533834227    0.680299575477431    0.154702508983149
                   AdaBoost   0.6851276348066045   0.3703960499168314   0.6119037461980517  0.23460639332618077
        3-Nearest-Neighbors  0.47121096592912115   0.3151179452688585   0.5289345881172548    0.294050924625357
        4-Nearest-Neighbors   0.5149354482196038   0.4023564611144884   0.5597845699144951   0.3732851264162055
        5-Nearest-Neighbors  0.48444422938347537  0.32197995387691625   0.5436262389808821    0.381503590924817
        6-Nearest-Neighbors   0.4233776115845601   0.3226385804494601   0.5011939908082478  0.25666124398180035
        7-Nearest-Neighbors  0.36612028345050845  0.27849056145729445   0.4306966040863781  0.17211850306430565
SVM (Radial Basis Function)  -0.8819853894116605  -0.9501091768184814  -2.3092099001076116   -2.055265235662462
              SVM (Sigmoid)  -0.8819853894116605  -0.9501091768184814  -2.3092099001076116   -2.055265235662462
               SVM (Linear)  -0.8819853894116605  -0.9501091768184814  -2.3092099001076116   -2.055265235662462
                 MLP (Relu)  -41.835878444440475   -67.52609741190454  -5.6783518244579145  -126.93479422973545
             MLP (Logistic)     -96.575585230635   -395.6920482278218  -36.797735882734585  -384.00241548442625
                 MLP (Tanh)   -55.28487635417859  -207.35191113699585  -11.853667133311912   -32.18861106178518
             MLP (Identity)  -17.830090517718226   -167.7636092197017  -39.629455079480664  -170.51152500274736
```

Table 19. 10-Fold Cross-Validated R^2 of Supervised Learning Algorithms

Ensemble methods (Random Forest and Adaboost) still performed best in predicting excess returns based on different factor weight allocations. Their cross-validated *R*^{2} scores are also diminished during contractionary period 4. SVM and MLP models fared poorly across all subperiods.

Large B/P	Large ROE	Large S/P	Large Return Rate in last quarter	Large Market Value	Small Systematic Risk	Predicted 20-year Excess Return
100%	0%	0%	0%	0%	0%	**1.110%**
0%	100%	0%	0%	0%	0%	**1.249%**
0%	0%	100%	0%	0%	0%	**1.701%**
0%	0%	0%	100%	0%	0%	**0.185%**
0%	0%	0%	0%	100%	0%	**0.223%**
0%	0%	0%	0%	0%	100%	**-0.078%**
16.67%	16.67%	16.67%	16.67%	16.67%	16.67%	**2.128%**

Table 20. Random Forest Estimate of 20-Year Excess Returns with various Factor Allocations

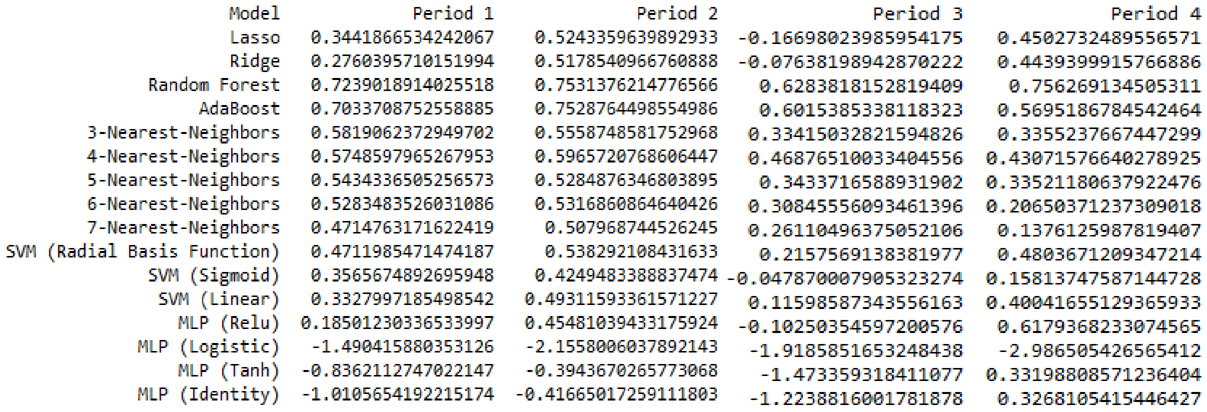
Random Forest algorithm predicted an equally weighted portfolio across the six factor-based strategies to generate the highest 20-year excess return among the 7 different scenarios. It estimated a portfolio 100% allocated to *Small Systematic Risk* strategy to yield the lowest 20-year excess return.

Large B/P	Large ROE	Large S/P	Large Return Rate in last quarter	Large Market Value	Small Systematic Risk	Predicted 20-year Excess Return
100%	0%	0%	0%	0%	0%	**1.206%**
0%	100%	0%	0%	0%	0%	**1.230%**
0%	0%	100%	0%	0%	0%	**1.633%**
0%	0%	0%	100%	0%	0%	**-0.025%**
0%	0%	0%	0%	100%	0%	**0.058%**
0%	0%	0%	0%	0%	100%	**-0.092%**
16.67%	16.67%	16.67%	16.67%	16.67%	16.67%	**1.959%**

Table 21. Adaboost Estimate of 20-Year Excess Returns with various Factor Allocations

Similar with Random Forest algorithm, Adaboost estimated an equally weighted portfolio across the six factor-based strategies to generate the highest 20-year excess return among the 7 scenarios. It predicted a portfolio 100% allocated to *Small Systematic Risk* strategy to yield the lowest 20-year excess return.

Building a Portfolio Risk Model on Systematic Risk



Model	Period 1	Period 2	Period 3	Period 4
Lasso	0.3441866534242067	0.5243359639892933	-0.16698023985954175	0.4502732489556571
Ridge	0.2760395710151994	0.5178540966760888	-0.07638198942870222	0.4439399915766886
Random Forest	0.7239018914025518	0.7531376214776566	0.6283818152819409	0.756269134505311
AdaBoost	0.7033708752558885	0.7528764498554986	0.6015385338118323	0.5695186784542464
3-Nearest-Neighbors	0.5819062372949702	0.5558748581752968	0.33415032821594826	0.3355237667447299
4-Nearest-Neighbors	0.5748597965267953	0.5965720768606447	0.46876510033404556	0.43071576640278925
5-Nearest-Neighbors	0.5434336505256573	0.5284876346803895	0.3433716588931902	0.33521180637922476
6-Nearest-Neighbors	0.5283483526031086	0.5316860864640426	0.30845556093461396	0.20650371237309018
7-Nearest-Neighbors	0.4714763171622419	0.507968744526245	0.26110496375052106	0.1376125987819407
SVM (Radial Basis Function)	0.4711985471474187	0.538292108431633	0.2157569138381977	0.4803671209347214
SVM (Sigmoid)	0.3565674892695948	0.4249483388837474	-0.047870007905323274	0.15813747587144728
SVM (Linear)	0.3327997185498542	0.49311593361571227	0.11598587343556163	0.40041655129365933
MLP (Relu)	0.18501230336533997	0.45481039433175924	-0.10250354597200576	0.6179368233074565
MLP (Logistic)	-1.490415880353126	-2.1558006037892143	-1.9185851653248438	-2.986505426565412
MLP (Tanh)	-0.8362112747022147	-0.3943670265773068	-1.473359318411077	0.33198808571236404
MLP (Identity)	-1.0105654192215174	-0.41665017259111803	-1.2238816001781878	0.3268105415446427

Table 22. 10-Fold Cross-Validated R^2 of Supervised Learning Algorithms

Ensemble methods (Random Forest and Adaboost) performed best in predicting systematic risk based on different factor weight allocations. Their cross-validated R^2 scores remained high even during the financial crisis in period 4. MLP fared badly and yielded negative R^2 scores in most subperiods.

Large B/P	Large ROE	Large S/P	Large Return Rate in last quarter	Large Market Value	Small Systematic Risk	Predicted Systematic Risk
100%	0%	0%	0%	0%	0%	**1.2337**
0%	100%	0%	0%	0%	0%	**1.1207**
0%	0%	100%	0%	0%	0%	**1.2285**
0%	0%	0%	100%	0%	0%	**1.2936**
0%	0%	0%	0%	100%	0%	**1.0518**
0%	0%	0%	0%	0%	100%	**1.0635**
16.67%	16.67%	16.67%	16.67%	16.67%	16.67%	**0.9265**

Table 23. Random Forest Estimate of Systematic Risk with various Factor Allocations

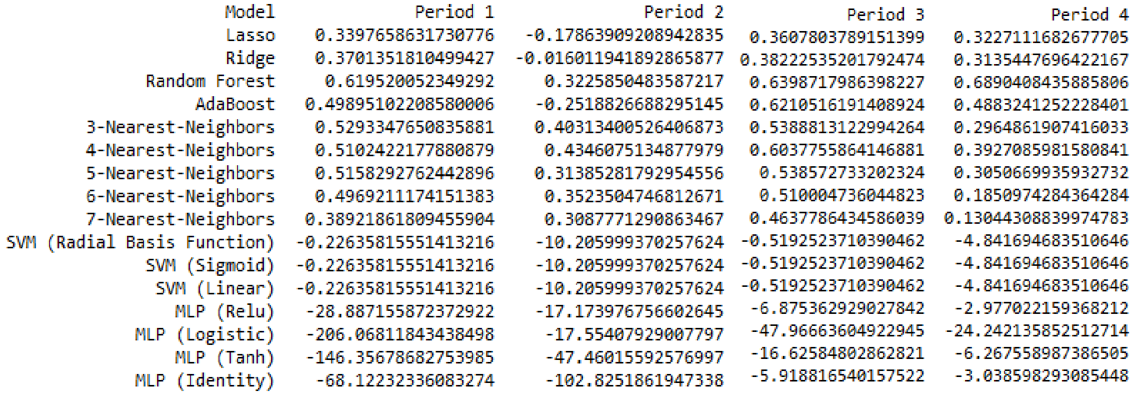
Random Forest algorithm predicted an equally weighted portfolio across the six factor-based strategies to generate the lowest systematic risk among the 7 given scenarios. It forecasted a portfolio 100% allocated to Momentum factor (*Large Return Rate in the last quarter*) to result to the highest systematic risk.

Large B/P	Large ROE	Large S/P	Large Return Rate in last quarter	Large Market Value	Small Systematic Risk	Predicted 20-year Systematic Risk
100%	0%	0%	0%	0%	0%	**1.2829**
0%	100%	0%	0%	0%	0%	**1.1000**
0%	0%	100%	0%	0%	0%	**1.2530**
0%	0%	0%	100%	0%	0%	**1.3900**
0%	0%	0%	0%	100%	0%	**1.0821**
0%	0%	0%	0%	0%	100%	**1.0405**
16.67%	16.67%	16.67%	16.67%	16.67%	16.67%	**0.9410**

Table 24. Adaboost Estimate of Systematic Risk with various Factor Allocations

Similar with Random Forest algorithm, Adaboost estimated an equally weighted portfolio across the six factor-based strategies to generate the lowest systematic risk among the 7 scenarios. It predicted a portfolio 100% allocated to Momentum factor (*Large Return Rate in the last quarter*) to yield the highest systematic risk.

Building a Portfolio Risk Model on Total Risk



```
                      Model              Period 1               Period 2              Period 3              Period 4
                      Lasso    0.3397658631730776   -0.17863909208942835    0.3607803789151399    0.3227111682677705
                      Ridge    0.3701351810499427  -0.016011941892865877   0.38222535201792474    0.3135447696422167
              Random Forest     0.619520052349292     0.3225850483587217    0.6398717986398227    0.6890408435885806
                   AdaBoost   0.49895102208580006    -0.2518826688295145    0.6210516191408924    0.4883241252228401
        3-Nearest-Neighbors    0.5293347650835881    0.40313400526406873    0.5388813122994264    0.2964861907416033
        4-Nearest-Neighbors    0.5102422177880879     0.4346075134877979    0.6037755864146881    0.3927085981580841
        5-Nearest-Neighbors    0.5158292762442896    0.31385281792954556     0.538572733202324    0.3050669935932732
        6-Nearest-Neighbors    0.4969211174151383     0.3523504746812671     0.510004736044823    0.1850974284364284
        7-Nearest-Neighbors   0.38921861809455904     0.3087771290863467    0.4637786434586039   0.13044308839974783
SVM (Radial Basis Function)  -0.22635815551413216    -10.205999370257624   -0.5192523710390462    -4.841694683510646
              SVM (Sigmoid)  -0.22635815551413216    -10.205999370257624   -0.5192523710390462    -4.841694683510646
               SVM (Linear)  -0.22635815551413216    -10.205999370257624   -0.5192523710390462    -4.841694683510646
                 MLP (Relu)   -28.887155872372922    -17.173976756602645    -6.875362929027842    -2.977022159368212
             MLP (Logistic)   -206.06811843438498     -17.55407929007797    -47.96663604922945   -24.242135852512714
                 MLP (Tanh)   -146.35678682753985     -47.46015592576997    -16.62584802862821    -6.267558987386505
             MLP (Identity)    -68.12232336083274     -102.8251861947338    -5.918816540157522    -3.038598293085448
```

Table 25. 10-Fold Cross-Validated R^2 of Supervised Learning Algorithms

Random Forest performed best in predicting total risk based on different factor weight allocations. Its cross-validated R^2 score remained high even during the financial crisis in period 4. SVM and MLP performed poorly and yielded negative R^2 scores in most subperiods.

Large B/P	Large ROE	Large S/P	Large Return Rate in last quarter	Large Market Value	Small Systematic Risk	Predicted 20-year Total Risk
100%	0%	0%	0%	0%	0%	**0.1337**
0%	100%	0%	0%	0%	0%	**0.1072**
0%	0%	100%	0%	0%	0%	**0.1370**
0%	0%	0%	100%	0%	0%	**0.1352**
0%	0%	0%	0%	100%	0%	**0.0938**
0%	0%	0%	0%	0%	100%	**0.1104**
16.67%	16.67%	16.67%	16.67%	16.67%	16.67%	**0.097**

Table 26. Random Forest Estimate of Total Risk with various Factor Allocations

Random Forest algorithm predicted an equally weighted portfolio across the six factor-based strategies to generate the lowest total risk of the 7 given scenarios. It forecasted a portfolio 100% allocated to Value factor (*Large S/P*) to return the highest total risk.

Implementation of supervised learning algorithms to build a model of portfolio return and risk reveals that ensemble methods (Random Forest and Adaboost) perform best in the given dataset. Moreover, using ensemble methods to predict portfolio returns shows that diversification across six factor-based strategies will maximize both annual and excess returns. Full exposure to *Small Systematic Risk* will yield lower annual and excess returns. Using the same ensemble algorithms also suggests that diversification across six factor-based strategies will lower systematic and total risk. 100% exposure to Momentum factor (*Large Return Rate in the last quarter*) and Value factor (*Large S/P*) may increase systematic and total risk, respectively.

8. Conclusion

This study determined the key features that have great mutual dependencies with each of the six performance indicators namely Annual Returns, Excess Returns, Systematic Risk, Total Risk, Absolute Win Rate, and Relative Win Rate (see *Table 10*). Gathering data on the specified features for each performance indicator can allow investors to monitor critical variables that drive the indicators.

This research also performed K-Means clustering to select the cluster of factor allocations that yielded the highest Sharpe Ratio or risk-adjusted return for each subperiod. Results show that forming a cluster of factor-based strategies with highest Sharpe Ratio in the current period can generate a higher-than-average Sharpe Ratio in the next 5-year period. Further analysis provides insight on the importance of a more diversified portfolio in the recent periods. It also implies that the risk and return profile of factor-based strategies become more similar over the years.

Implementation of supervised learning algorithms also shows that ensemble methods (Random Forest and Adaboost) prove to perform better than other algorithms (K-Nearest-Neighbor, Support Vector Machine, Multi-Layer Perceptrons, Lasso and Ridge Regression) in building a reliable model that links factor weight allocations to different measures of portfolio risks and returns. Moreover, adopting a diversified exposure across the six factor-based strategies can maximize long-term returns and minimize risks. 100% allocation to small systematic risk may minimize long-term returns. 100% allocation to Momentum factor and Value factor may be suboptimal in reducing the level of portfolio risk.

9. References

Arnott, R, Harvey C, Kalesnik V, and Linnainmaa J (2019) “Alice’s Adventures in Factorland: Three Blunders That Plague Factor Investing.” Journal of Portfolio Management, vol. 45, no. 4: 18–36

Banz RW (1981) The relationship between return and market value of common stocks. J Financ Econ 9:3–18

Fama E, French K (1998) Value versus growth: the international evidence. J Financ 53(6):1975–2000

Harvey, C., Liu Y, and H. Zhu (2016) “… and the CrossSection of Expected Returns.” Review of Financial Studies 29, no. 1 (January): 5–68.

Jegadeesh N, Titman S (1993) Returns to buying winners and selling losers: implications for stock market efficiency. J Financ 48(1):65–91

Liu, Y, Yeh, C. (2015). Using mixture design and neural networks to build stock selection decision support systems. Neural Computing and Applications. 28. 10.1007/s00521-015-2090-x.