Factor Investing Strategies Based on Specific Feature and Random Forest Method

Name: Yaxin LI

Contents

1. Introduction

In this report, my goal is building factor investing strategies based on a specific feature and random forest method, and compare the performance of different strategies with the benchmark, which is the equal weighted portfolio with all the stocks.

First of all, I prepare the whole dataset for further study. Here, a data analysis with descriptive statistics is given so as to help us get a full picture of the dataset. Then, I normalize the data so as to make it suitable for further use. In addition, the autocorrelation is also analyzed. Considering that compared with the high autocorrelation of eleven features of stocks, the future returns have little autocorrelation, I will also use the variation of features for building machine learning based strategies.

Secondly, for the factor investing strategy based on a specific feature, a proper feature with a proper threshold should be chosen so as to build the portfolio. Thus, a grid research is carried out here to achieve the goal. Furthermore, to see if the performance of the portfolio is significant better compared with other portfolios, I also carry out a deflated Sharpe ratio test in this study.

Thirdly, I build two factor investing strategies based on random forest method, one with the features while the other with the deviation of features.

Last but not least, the performance of the three strategies mentioned above will be compared with the benchmark, and plenty of performance measures will be used to compare the performance.

2. Data Preparation

2.1 Data Analysis

This dataset consists of monthly financial data of 209 stocks from 1999-02-01 to 2008-04-03. Eleven features are provided in the dataset: market capitalization (Mkt_Cap), price to book ratio (P2B), Volume of 1 month (Vol_1M), dividend yield (Div_yield), PE ratio (PE_ratio), RSI index of 1 month (RSI_1M), debt to equity ratio (D2E), profit growth (Prof_growth), difference of return of different caps (Ret_Cap), asset growth (Asset_growth) and profit margin (Prof_Marg). Table 2-2 shows the first 6 rows of the dataset.

In addition, the future returns are calculated here by using the following formula:

$$F\_Return_t = \frac{Close\ price_{t+1}}{Close\ price_t} - 1 \qquad (2-1)$$

Table 2-1 First 6 rows of the dataset

Tick <fctr>	Date <date>	Close <dbl>	Mkt_Cap <dbl>	P2B <dbl>	Vol_1M <dbl>	Div_yield <dbl>	PE_ratio <dbl>	RSI_1M <dbl>	D2E <dbl>
AA	1999-02-01	46.4358	30422.231	1.8907	45.314	2.4227	12.6874	55.6823	46.3243
AAPL	1999-02-01	1.4621	5534.463	3.1251	60.727	0.5749	17.4947	53.6500	49.6100
ABT	1999-02-01	20.5029	69573.428	5.4545	32.109	2.9264	13.6686	48.0191	53.8592
ABX	1999-02-01	19.6875	7402.500	2.0663	39.126	0.9143	24.9209	49.8288	13.9198
ADBE	1999-02-01	5.9844	2915.157	5.6455	43.491	0.4178	25.3307	53.9088	0.0000
ADM	1999-02-01	13.3895	9195.344	1.4048	37.268	1.2442	23.3358	44.0353	67.6810

Vol_1M <dbl>	Div_yield <dbl>	PE_ratio <dbl>	RSI_1M <dbl>	D2E <dbl>	Prof_growth <dbl>	Ret_Cap <dbl>	Asset_growth <dbl>	Prof_Marg <dbl>	F_Return <dbl>
45.314	2.4227	12.6874	55.6823	46.3243	25.7427	13.0189	33.6014	5.1991	-0.002261186
60.727	0.5749	17.4947	53.6500	49.6100	36.5439	5.9460	11.2942	8.8889	-0.175569386
32.109	2.9264	13.6686	48.0191	53.8592	2.0639	29.4837	9.9398	18.7925	0.002726444
39.126	0.9143	24.9209	49.8288	13.9198	7.2581	7.5761	8.1050	22.5543	-0.104761905
43.491	0.4178	25.3307	53.9088	0.0000	9.4071	17.0718	-18.3752	20.3732	-0.160584186
37.268	1.2442	23.3358	44.0353	67.6810	16.2743	5.1481	4.9116	2.4315	-0.004033011

To get more detail, some descriptive statistics are given as follows.

Table 2-2 Descriptive statistics of data

Statistics	Mkt_Cap	P2B	Vol_1M	Div_yield	PE_ratio	RSI_1M
Min	17.9	0.059	1.138	0.0013	0.374	19.70
Median	10855.3	2.619	27.106	1.8061	17.311	51.99
Mean	29712.9	6.872	32.083	2.2023	31.559	51.76
Max	887952.3	3304.600	485.029	90.1292	6626.039	84.84
Statistics	D2E	Prof_growth	Ret_Cap	Asset_growth	Prof_Marg	F_Return
Min	0.00	-1141.139	-267.57	-80.0846	-670.053	19.70
Median	64.38	5.924	11.34	5.1193	7.188	51.99
Mean	278.69	19.138	12.24	9.7056	6.972	51.76
Max	130233.33	21750	927.92	763.8937	228.987	84.84

We can see from Table 2-2 that for most of the features, the differences between minimum, median and maximum are indeed huge, which indicate that there are some outliers in the dataset. For instance, the maximum of P2B is 491.756 times larger than its median, and its median is 116.475 times larger than its minimum. The situation is similar to PE ratio, D2E, etc. This is partly because sometimes we will get a really small number in the denominator to calculate the ratio.

Apart from the huge dispersion inside each feature, the difference of the absolute values of different features are also huge. Thus, the data of 11 features need to be normalized before we do some further study.

2.2 Normalizing the Data

There are several methods to normalize data. However, the max & min method will suffer more on the huge outliers, sometimes leading to lack of data in the middle of the interval, and the mean & standard deviation method is more suitable for data which is normally distributed, and the data may not in the interval [-1,1]. Thus, we use the empirical cumulative density function to normalize each feature to the interval [0,1]. By using this method, the information of the ranks of the data will be kept, and the data will be homogeneous. The formula I used is as follows.

$$\widetilde{Feature}_k = F_k(\widetilde{Feature}_k) \qquad (2-2)$$

where the $F_k(\widetilde{Feature}_k)$ is the empirical cdf of the k^{th} feature.

We can get the following data after normalization. From Table 2-3, all the data of features are indeed homogenous now and in the interval [0,1].

Table 2-3 Features after normalization

Tick <fctr>	Date <date>	Close <dbl>	Mkt_Cap <dbl>	P2B <dbl>	Vol_1M <dbl>	Div_yield <dbl>	PE_ratio <dbl>	RSI_1M <dbl>	D2E <dbl>
AA	1999-03-01	46.3308	0.7368421	0.3301435	0.3492823	0.6937799	0.2535885	0.7655502	0.33971292
AAPL	1999-03-01	1.2054	0.4545455	0.4736842	0.8229665	0.1722488	0.3444976	0.2679426	0.36842105
ABT	1999-03-01	20.5588	0.9282297	0.7799043	0.4784689	0.7607656	0.3205742	0.6028708	0.41148325
ABX	1999-03-01	17.6250	0.5454545	0.3157895	0.6411483	0.3062201	0.5837321	0.2775120	0.09569378
ADBE	1999-03-01	5.0234	0.2966507	0.7177033	0.8038278	0.1531100	0.5454545	0.2966507	0.02392344
ADM	1999-03-01	13.3355	0.6507177	0.1818182	0.2392344	0.3732057	0.6267943	0.3971292	0.51196172

Prof_growth <dbl>	Ret_Cap <dbl>	Asset_growth <dbl>	Prof_Marg <dbl>	F_Return <dbl>
0.7990431	0.5885167	0.88516746	0.4497608	-0.01208483
0.8612440	0.1483254	0.60287081	0.6555024	0.06844201
0.4210526	0.9473684	0.55502392	0.9330144	0.03537658
0.5311005	0.2631579	0.49760766	0.9569378	-0.04255319
0.5741627	0.7703349	0.02392344	0.9377990	0.40281483
0.7177033	0.1244019	0.37320574	0.2488038	-0.06477447

Moreover, if we draw the histogram of the rescaled features, we can see from Figure 2-1 that now the feature are almost uniformly distributed during the interval [0,1].

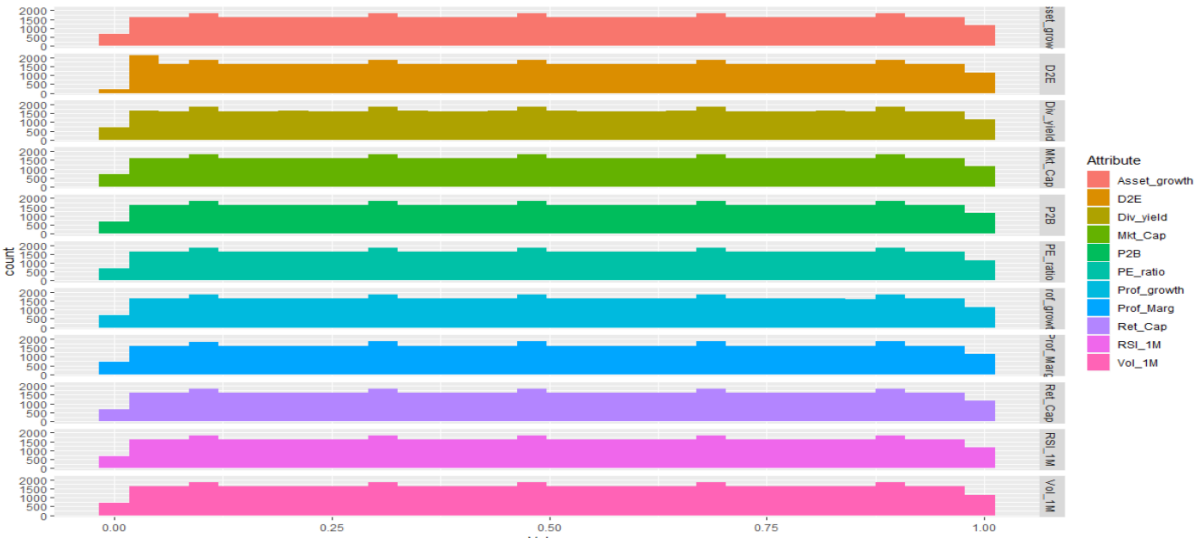

Figure 2-1 Sharpe ratios of different portfolios

2.3 Autocorrelation of the Data

Considering features may be highly correlated, I calculated the first order autocorrelation coefficients of each features and future returns, and the results for the first 10 stocks are as follows.

Table 2-4 Autocorrelation for the first 10 stock

Tick <fctr>	Close <dbl>	Mkt_Cap <dbl>	P2B <dbl>	Vol_1M <dbl>	Div_yield <dbl>	PE_ratio <dbl>	RSI_1M <dbl>	D2E <dbl>	Prof_growth <dbl>
AA	0.9708603	0.9799849	0.9709620	0.6394011	0.9750454	0.9161613	0.4826379	0.9204820	0.8905276
AAPL	0.9799852	0.9832863	0.9735048	0.6859572	0.9749120	0.9659044	0.5152700	0.9582224	0.9277482
ABT	0.9670853	0.9631740	0.9395027	0.5281348	0.9753913	0.9796735	0.4602194	0.9450579	0.8682419
ABX	0.9619298	0.9612896	0.9158972	0.5978907	0.9390105	0.9489434	0.2790646	0.9789923	0.8666869
ADBE	0.9476322	0.9253708	0.9575464	0.4940243	0.9267743	0.8865520	0.4655200	0.9841640	0.8765700
ADM	0.9771760	0.9050808	0.9124281	0.6258531	0.9194366	0.9220751	0.4793051	0.9782315	0.7916201
ADP	0.9742953	0.9652121	0.9460350	0.4787216	0.9786922	0.9506828	0.3349779	0.8960729	0.8717372
AEP	0.9692332	0.9321306	0.8503620	0.5952611	0.9233603	0.8868001	0.4237559	0.9771060	0.8864771
AES	0.9765674	0.9717739	0.9581636	0.6271932	0.9795210	0.9282739	0.5818170	0.9734797	0.7592027
AGCO	0.9684769	0.9401389	0.9360405	0.4706994	0.9759084	0.9684127	0.4167773	0.9756562	0.9308810

Div_yield <dbl>	PE_ratio <dbl>	RSI_1M <dbl>	D2E <dbl>	Prof_growth <dbl>	Ret_Cap <dbl>	Asset_growth <dbl>	Prof_Marg <dbl>	F_Return <dbl>
0.9750454	0.9161613	0.4826379	0.9204820	0.8905276	0.9728660	0.8317093	0.8819341	-1.671983e-02
0.9749120	0.9659044	0.5152700	0.9582224	0.9277482	0.9755429	0.9544572	0.9728653	5.392373e-03
0.9753913	0.9796735	0.4602194	0.9450579	0.8682419	0.9416614	0.9046099	0.7354834	-5.068712e-02
0.9390105	0.9489434	0.2790646	0.9789923	0.8666869	0.9461630	0.9442995	0.7875020	-1.573703e-01
0.9267743	0.8865520	0.4655200	0.9841640	0.8765700	0.9863217	0.9190791	0.8949736	3.387718e-02
0.9194366	0.9220751	0.4793051	0.9782315	0.7916201	0.9414526	0.9330840	0.7639165	5.397052e-02
0.9786922	0.9506828	0.3349779	0.8960729	0.8717372	0.9195759	0.7832799	0.7134074	-5.423890e-02
0.9233603	0.8868001	0.4237559	0.9771060	0.8864771	0.9119019	0.9239491	0.6596242	4.634197e-02
0.9795210	0.9282739	0.5818170	0.9734797	0.7592027	0.9377373	0.9398724	0.7553037	7.493376e-02
0.9759084	0.9684127	0.4167773	0.9756562	0.9308810	0.9823549	0.9458109	0.8256164	-4.759206e-02

We can see from Table 2-4 that the features are highly autocorrelated, while the future returns are almost not autocorrelated. Thus, there is another option for us, and we can also use the deviation of the features to build the machine learning based model. Table 2-5 shows the autocorrelation coefficients of the deviations of the features, and we can see that they are just mildly autocorrelated.

Table 2-5 Autocorrelation of the deviation of the features

Tick <fctr>	Date <date>	Mkt_Cap_varia <dbl>	P2B_varia <dbl>	Vol_1M_varia <dbl>	Div_yield_varia <dbl>	PE_ratio_varia <dbl>	RSI_1M_varia <dbl>
AA	1999-03-01	-0.119617225	0.009569378	-0.33971292	-0.009569378	0.01913876	0.03349282
AAPL	1999-03-01	-0.047846890	-0.086124402	-0.09090909	-0.014354067	-0.09569378	-0.37320574
ABT	1999-03-01	0.004784689	0.023923445	0.15789474	-0.023923445	0.03349282	0.18660287
ABX	1999-03-01	-0.014354067	-0.038277512	0.09090909	0.014354067	-0.04784689	-0.19617225
ADBE	1999-03-01	-0.028708134	-0.062200957	0.15311005	0.028708134	-0.09569378	-0.35406699
ADM	1999-03-01	0.000000000	0.028708134	-0.25837321	-0.023923445	0.02870813	0.12918660

D2E_varia <dbl>	Prof_growth_varia <dbl>	Ret_Cap_varia <dbl>	Asset_growth_varia <dbl>	Prof_Marg_varia <dbl>
0.000000000	0.000000000	0	-0.014354067	-0.004784689
0.000000000	0.000000000	0	-0.004784689	-0.009569378
-0.004784689	0.000000000	0	-0.004784689	0.000000000
0.000000000	-0.004784689	0	0.000000000	0.004784689
0.004784689	-0.004784689	0	0.000000000	0.000000000
0.000000000	0.000000000	0	0.000000000	0.000000000

Now, we are ready to use the prepared data to build our strategies.

3. Strategies Based on a Specific Feature

3.1 Grid Research for the Eleven Features

Since there are 11 features in the dataset and we just want to use one feature to build our portfolio, we need to find a proper feature with a proper threshold which could give us the best performance. Thus, a grid research is carried out here to achieve the goal. In the grid research, for each feature, seven thresholds which are 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 2 directions which are 1 and -1 are used. In other word, we will build 7*2=14 buy-only portfolios for each feature, and for each threshold, both the stocks with feature values above the threshold and the stocks with feature values below the threshold will be used to build portfolios respectively. Thus, we have 14*11=154 portfolios in total, and Shape ratio are calculated here as the performance measures. Here, we suppose the risk-free rate is 0, so the formula for Sharpe ratio can be simplified as follows.

$$Sharpe\ ratio = \frac{\bar{r}}{\sigma} \qquad (3-1)$$

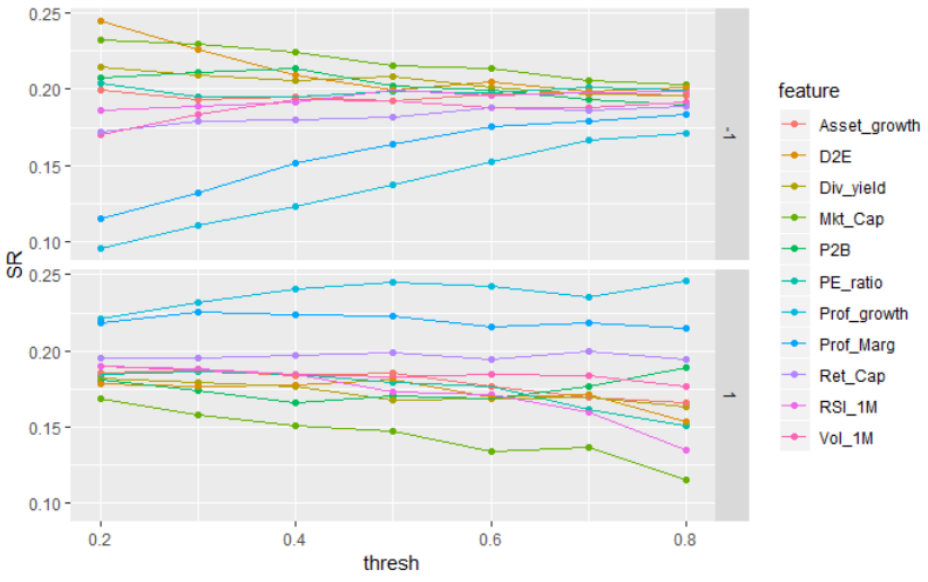

Figure 3-1 Sharpe ratios of different portfolios

We can draw the conclusion from Figure 3-1 that the portfolio with profit growth larger than 0.8 have the best performance among all the portfolios. This is reasonable because the profit growth is a measure of the growth rate of the profit. As is often the case, the companies with high profit growth usually have good developing prospects, thus leading to the increase of the stock prices and high Sharpe ratios.

In addition, the portfolio with debt to equity ratio smaller than 0.2 ranks second among all the portfolios. This is also reasonable because low debt to equity ratio means low leverage, and this usually makes the companies be more stable, thus contributing to relatively low volatility and high Shape ratio.

3.2 Deflated Sharpe Ratio Test

However, the Sharpe ratio might be overestimated, and we can carry out a deflated Sharpe ratio test to test the significance of the best portfolio we have found. The deflated Sharpe ratio test is based on the idea that we can compare the best Sharpe ratio obtained in the back test to a theoretical value for the average maximum Sharpe ratio and test whether the difference is significant. If the statistic is above a certain threshold (e.g. 0.95), the Sharpe ratio is significant, otherwise it is not significant.

The formula of the statistic is as follows:

$$t = \phi\left(\frac{(SR - SR^*) * \sqrt{T-1}}{\sqrt{1 - \gamma_3 SR + \frac{\gamma_4 - 1}{4} SR^2}}\right) \quad (3-2)$$

where SR is the Sharpe ratio obtained by the best strategy, and SR^* is the theoretical average best Sharpe ratio. The formula for SR^* is as follows.

$$SR^* = E(SR_m) + \sqrt{V(SR_m)}\left((1-\gamma)\phi^{-1}\left(1 - \frac{1}{N}\right) + \gamma\phi^{-1}\left(1 - \frac{1}{Ne}\right)\right) \quad (3-3)$$

In addition, γ_3 and γ_4 are the skewness and kurtosis of the returns of the chosen strategy; ϕ is the cumulative density function (cdf) of the standard Gaussian law and γ is the Euler-Mascheroni constant; the index *N* refers to the number of strategy we tried which are 154 here, and $E(SR_m)$ and $V(SR_m)$ are the mean and variance of all the portfolios.

The deflated Sharpe ratio test (DSR test) is carried out for the portfolio with profit growth larger than 0.8, and the result is shown in Table 3-1.

Table 3-1 Result of DSR test

Strategies	t value
Portfolio with profit growth larger than 0.8	0.4294

In Table 3-1, the t value of the portfolio is far from the threshold value (e.g. 0.95), which means that although the portfolio seems to have the best performance according to Sharpe ratio among all the portfolios, it actually does not have significant better performance.

However, since the profit growth with threshold 0.8 is the best portfolio we could find up to now, still, I will build a portfolio with profit growth larger than 0.8 and compare its performance with the benchmark.

4. Factor Investing Strategies Based on Random Forest Model

4.1 Presentation of Simple Tree Model

In some cases, the relationship between the dependent variable and independent variable is not linear, and we can use the tree model to classify the dependent variable, which is the future returns in our case, according to different values of independent variables, which are the features and the deviation of the features in our case. The basic

idea is that in each node, we should split the data into two group according to a specific feature with a specific value which could help us get homogenous clusters. In other word, if we use total quadratic variation $\mathrm{V(c)}$, we should choose the feature and threshold which can give us the minimum value of the sum of the dispersion inside groups. And the minimization program is shown as follows.

$$\mathrm{V(c)} = \min[\sum_{i:feature_k>c} (y - \bar{y}_{feature_k>c})^2 + \sum_{i:feature_k<c} (y - \bar{y}_{feature_k<c})^2] \qquad (4-1)$$

where y is the dependent variable.

Furthermore, if we use cp = 0.001 and maximum depth = 3 with the whole dataset to train the tree model, we will get Figure 4-1.

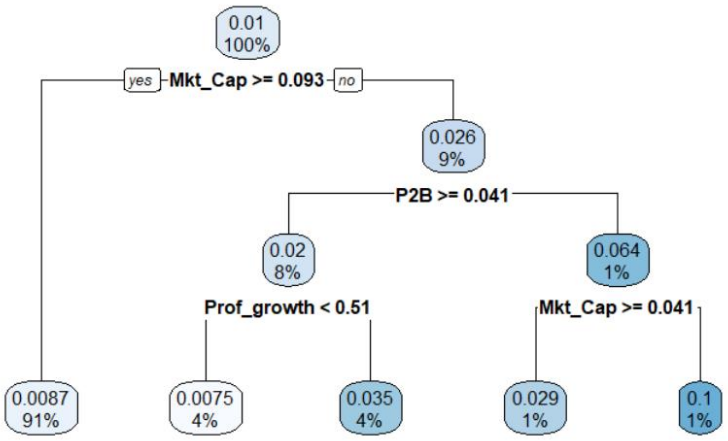

Figure 4-1 Simple tree model with cp = 0.001 and maximum depth = 3

As we can see from Figure 4-1, the split is not that satisfying because 91% of the data falls into the first group. This situation wouldn't get better even if we use maximum depth = 4, as is shown in Figure 4-2.

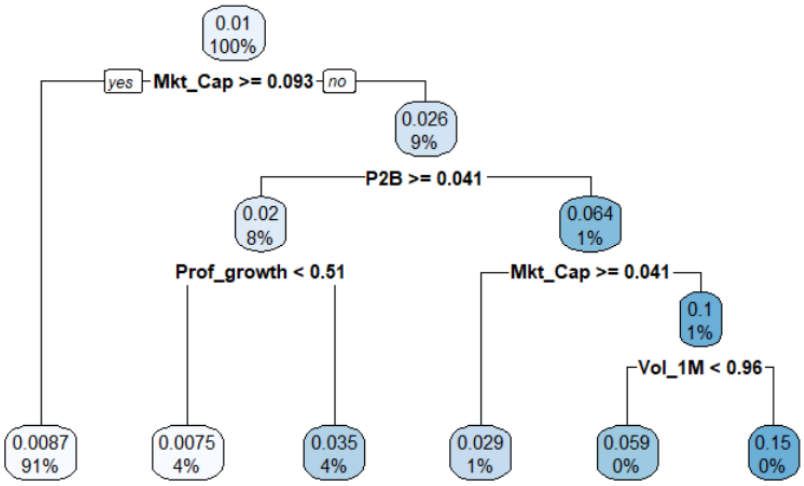

Figure 4-2 Simple tree model with cp = 0.001 and maximum depth = 4

4.2 Strategies Based on Random Forest Model

In fact, we can only get one optimal tree with a splitting rule and a dataset. What's more, instead of using simple tree model, we can consider its revised version, which is the random forest method.

The random forest method uses the "bagging" technique. That is, instead of using all the features in the splitting process, each time we choose a subset of the features randomly and use them to build the tree. Then we aggregate all the trees so as to get the final tree. As a result, each final output value is the average value obtained over all the trees.

4.2.1 Essential Parameters for Random Forest Model

In this report, I use 7 out of 11 features to build the tree, and 10 random trees are generated each time so as to get the final tree.

Furthermore, in order to build the portfolio, we need to decide the weight of each stock at each period. Considering that the random forest method need some data to train the model, the data from 1999-03-01 to 2007-05-01 will only be used for the training purpose, and the end of the date will vary from 2007-06-01 to 2018-03-01. In addition, expanding window are used here to train the model. These essential parameters are summarized in Table 4-1.

Table 4-1 Essential parameters for random forest model

Number of features used each time	Number of trees generated each time	Start of the training period	End of the training period	Window
7	10	1999-03-01	[2007-06-01, 2018-03-01]	Expanding window

4.2.2 Transforming the Signal and Build the Portfolio

First of all, I transform the future returns into ordinal returns according to a specific return threshold, and get the new column denoted FC_Return. Different from future returns, the future constant returns only contain 3 values: 0, which indicates that the return is low; 1, which indicates the return is normal; 2, which indicates the return is high. Here, I take the threshold r as 0.02, and the future returns are transformed by using the following formula:

$$FC\_Return_t^i = \begin{cases} 0 & if\ F\_Return_t^i < -r \\ 1 & if\ \ F\_Return_t^i \in [-r, r] \\ 2 & if\ \ F\_Return_t^i > r \end{cases} \qquad (4-2)$$

Second, to decide the weight of each stock at each period, we need to transfer the signals generated from the random forest model into investment signals. I use the previous data to train the model, and then I use the model to predict the returns in the next period, which is exactly the values of FC_Return. In the next step, I rank the predicted values of FC_Return from the biggest to the smallest, choose the 30 stocks with the biggest predicted values of FC_Return and give them equal weights to build the portfolio.

In addition, considering that the features are highly autocorrelated while the future returns are almost not autocorrelated, two strategies based on random forest method will be built later using the features and the deviation of the features respectively. By doing this, we can compare the difference of the performance of the portfolios using predictors with different autocorrelation features.

5. Performance Comparison among All the Strategies

5.1 Cumulative Return and Maximum Drawdown

To compare the cumulative values of the four portfolios, we can draw the four evolution paths of the cumulative values in one graph, which is Figure 5-1. And we can draw the conclusion from Figure 5-1 that, during the financial crisis from year 2008 to 2009, all the portfolio performs badly, and we could loss half of the value of our portfolio. This can be explained by the fact that during the crisis, all the stocks perform badly and have decreasing prices; since we build four buy-only portfolios, it's normal that we will suffer from the crisis.

Furthermore, the second portfolio, which consists of stocks with the normalized profit growth higher than 0.8, achieves the highest value at the end of the date. The third portfolio, which consists of 30 stocks with the highest predicted value of FC_Returns based on random forest method with features, has somehow comparable performance with the second portfolio during 2007-06-01 to 2016-02-01, but performs worse than the second portfolio since 2016-02-01, and achieves the second large cumulative value portfolio at the end of the date. In addition, the forth portfolio, which consists of 30 stocks with the highest predicted value of FC_Returns based on random forest method with the deviation of features, has comparable performance with the benchmark, which is the equally weighted portfolio of all the stocks.

Last but not least, the second portfolio seems to be the most volatile one, then comes the third portfolio and the forth portfolio.

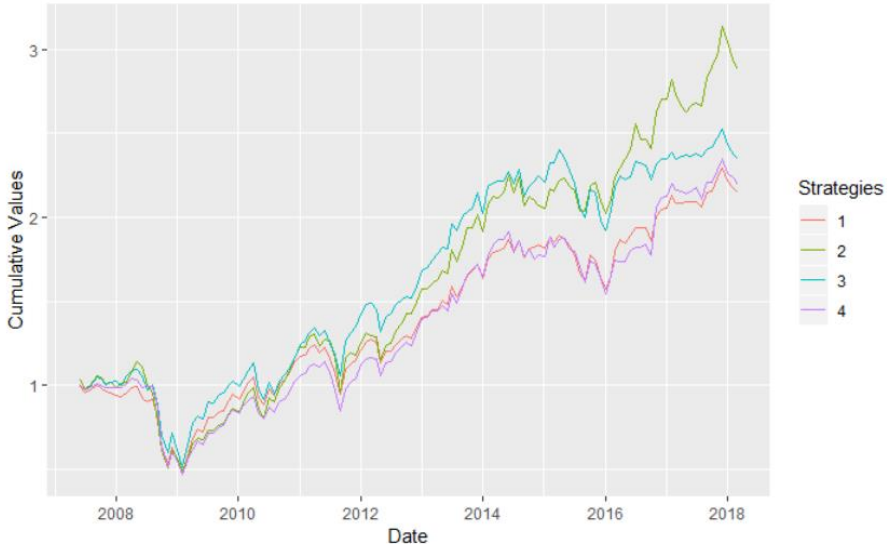

Figure 5-1 Cumulative value of all the portfolios

In addition, we also care about the maximum drawdown during the evolution of the cumulative values of our portfolios. The maximum drawdown (MDD) can be defined as the maximum decline from the historical peak. This ratio describes the maximum loss investor would suffer is the investor buy this portfolio at the time when the portfolio is at the historical peak, and the formula are as follows.

$$MDD = \max_{t\in[0,T]}\left(\left(\max_{\tau\in[0,t]} P_\tau\right) - P_t\right) \qquad (5-1)$$

Figure 5-2 shows the evolution of the cumulative values and maximum drawdowns of the four portfolios respectively. We can see that during the crisis, all the portfolios suffer a lot and the maximum drawdowns all appear during this period, which is consistent with the conclusion we get from Figure 5-1.

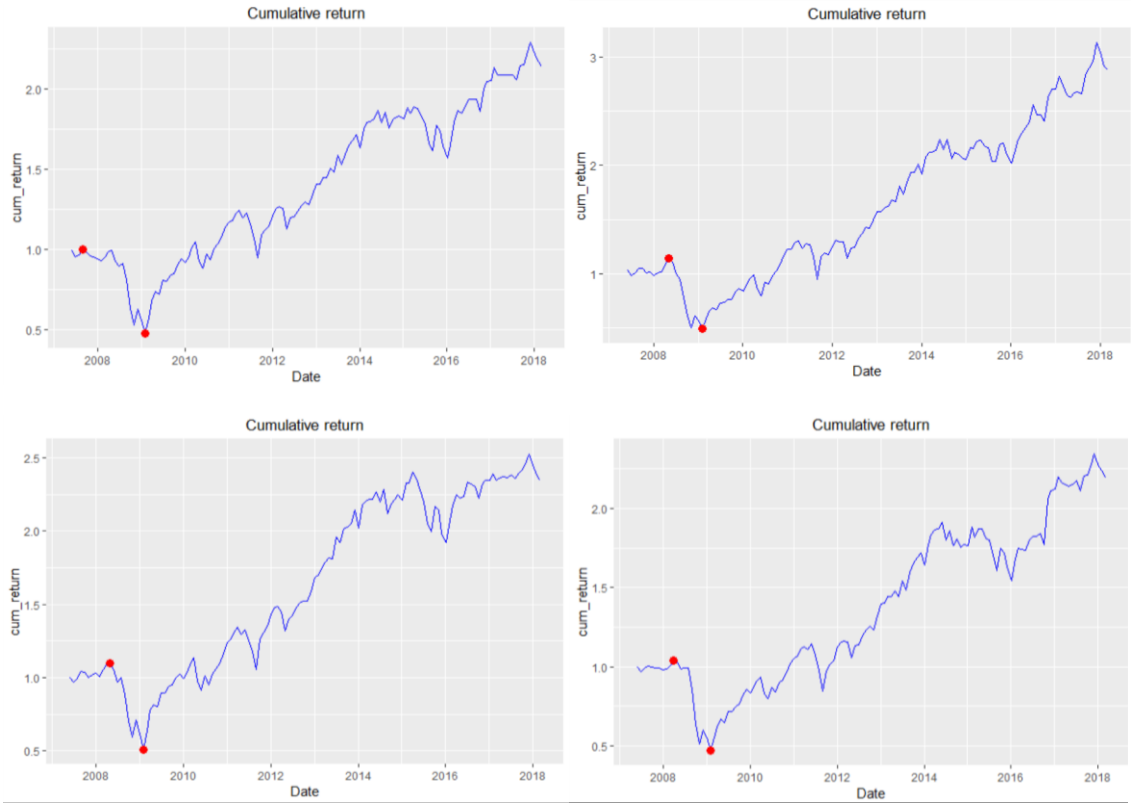

Figure 5-2 cumulative values and maximum drawdowns of 4 portfolios (the graphs at top left, top right, bottom left and bottom right are the graphs for the first, second, third and fourth portfolio respectively)

To be more precise, the exact values of the maximum drawdown are listed in Table 5-1. We can see that the second portfolio has the largest maximum drawdown 64.948%, which partly results from one of the disadvantage we already discovered of this portfolio: the high volatility. The high volatility may make the portfolio be able to gain a lot when the market conditions are fine, but will also make the portfolio suffer more when the market has a crisis.

Table 5-1 Maximum drawdown of all the strategies

Strategies	maximum drawdown
Benchmark (EW)	52.250%
Profit Growth	64.948%
RF with predictors	58.579%
RF with deviation of predictors	56.995%

5.2 Quantitative Performance Measures

Here, in order to compare the performance among all the strategies, lots of performance measures are calculated. To put it in the first place, several common performance measures are calculated, and they are: average return, volatility, Sharpe ratio, Value at Risk (VaR) at 95% confidence level and turnover rate. In addition, to make the results more realistic, the turnover-adjusted Sharpe ratio are also calculated here so as to take the transaction cost into account. Considering that we get all the returns based on monthly data, we should annualize all the performance measures, and the risk-free rate is supposed to be 0 as well.

The formulas I use to calculate these performance are as follows.

$$Average\ return = \overline{portfolio\ returns} * 12 \quad (5-2)$$

$$Volatility = var(portfolio\ returns) * \sqrt{12} \quad (5-3)$$

$$Sharpe\ ratio = \frac{Average\ return}{Volatility} \quad (5-4)$$

$$VaR95\%: \ P(portfolio\ returns_t * 12 < VaR95\%) = 5\% \quad (5-5)$$

For the turnover rate, we have the following formula:

$$Turnover\ rate = \frac{1}{T}\sum_{t=2}^{T}\sum_{n=1}^{N}|w_t^n - w_{t-}^n| * 12 \quad (5-6)$$

where $t-$ is the time just before rebalancing, and we should use the asset of the returns to calculate the trajectory of the weights.

In addition, since asset rotation is very costly, the turnover-adjusted Sharpe ratio are also calculated here so as to take the transaction cost into account. The formula of TC-adjusted Sharpe ratio is as follows. Here, I take the TC constant δ as 0.02.

$$TC - adjusted\ Sharpe\ ratio = \frac{Average\ return - \delta * Turnover\ rate}{Volatility} \quad (5-7)$$

What's more, since we already get the maximum drawdown (MDD) of each portfolio, we can calculate the MAR ratio which uses MDD to adjust the average return by using the following formula:

$$\text{MAR ratio} = \frac{Average\ return}{MDD} \quad (5-8)$$

The results of the above performance measures for the 4 models are as follows.

Table 5-2 Some performance measures of the four portfolios

Strategies	Average return	Volatility	Sharpe ratio	VaR 95%	Turnover rate	Turnover-adjusted Sharpe ratio	MAR ratio
Benchmark (EW)	0.093	0.210	0.441	-1.261	0.295	0.413	0.177
Profit Growth	0.125	0.232	0.540	-1.282	0.948	0.458	0.193
RF with features	0.104	0.223	0.464	-1.175	11.268	-0.547	0.177
RF with deviation of features	0.096	0.214	0.449	-1.087	11.091	-0.587	0.169

We can tell from Table 5-2 that according to Sharpe ratio, the second portfolio has the best performance, then comes the third portfolio. However, the second and third portfolio are more volatile than the first portfolio (benchmark), indicating that they are riskier as well. In addition, the forth portfolio has nearly the same performance with the benchmark according to average return, volatility and Sharpe ratio, which is consistent with the conclusion we draw from Figure 5-1.

Besides, if we take a look at the Value at Risk (VaR) at 95% confidence level, the second portfolio are the riskiest one with the largest absolute value of VaR, while the VaR of the third and fourth portfolio are smaller than the benchmark.

What's more, if we compute the turnover rate, we will discover huge differences among all the portfolios. The

turnover rate for the benchmark is 0.295, which indicates that 29.5% of our holding asset will be rotated over a year, while for the second portfolio is 0.948, which is 3.21 times larger than that of benchmark. As for the portfolios based on random forest method, the turnover rate is extremely big and it's about 37.6 times larger than that of the benchmark, and our holding asset will be rotated about 11 times in a year! That will definitely lead to high transaction cost, and it's not very feasible in the reality.

Furthermore, the TC-adjusted Sharpe ratios of the last two portfolio are negative even if the TC constant δ is just 0.02. Besides, the TC-adjusted Sharpe ratio of the second portfolio is 0.458, still higher than the value of benchmark which is 0.413. But we should note that if we put more penalty on the turnover ratio and use higher TC constant δ, the difference between the TC-adjusted Sharpe ratios of benchmark and the second portfolio will decrease. In particular, when δ is larger than 0.152, the benchmark will outperform the second portfolio and has the highest TC-adjusted Sharpe ratio.

Last but not least, we can tell from the four MAR ratios that although the second portfolio has the largest maximum drawdown, its MAR ratio is still the highest. The third portfolio has the same MAR ratio with benchmark, meaning that after adjusting by using maximum drawdown, its performance is somehow comparable with the benchmark. And the forth portfolio has the lowest MAR ratio.

6. Conclusions

In this report, three factor investing strategies based on a profit growth, random forest method with features, and random forest method with the deviation of features are built and their performance are compared with the benchmark, which is the equal weighted portfolio with all the stocks.

According to plenty of performance measures, the second portfolio performs the best according to Sharpe ratio, turnover-adjusted Sharpe ratio and MAR ratio, but it's also the most volatile portfolio which has the highest volatility, maximum drawdown and value at risk at 95% confidence level. The two strategies based on random forest method have really high turnover rates and have negative turnover-adjusted Sharpe ratios, making them be costly in the reality. In addition, the portfolio based on random forest method with features have the same MAR ratio compared with the benchmark. Last but not least, the portfolio based on random forest method has the worst performance among the three factor investing strategies according to Sharpe ratio, turnover-adjusted Sharpe ratio and MAR ratio.