Report

The trading problem of the project was to achieve an optimal in-sample and out-of-sample cumulative return during a given date range for a specific stock symbol considering the trading costs. To turn it into a learning problem, I devised a plan to set up 20 random tree learners, using a bag learner. For the random tree learner, I had to turn the regression learner into classification learner so that once the values were fed to the learner, it would only pick from whatever Y values were fed to the learner and not pick the mean of some values from the Y data. To make the transformation from the regression learner to the classification learner, I took the mode of the Y data where I was taking the mean. For the leaf size, I decided to use a value greater than 5 to avoid degenerate overfitting for in-sample data. It turned out that bag size 30 and leaf size 5 worked out to produce the best results.

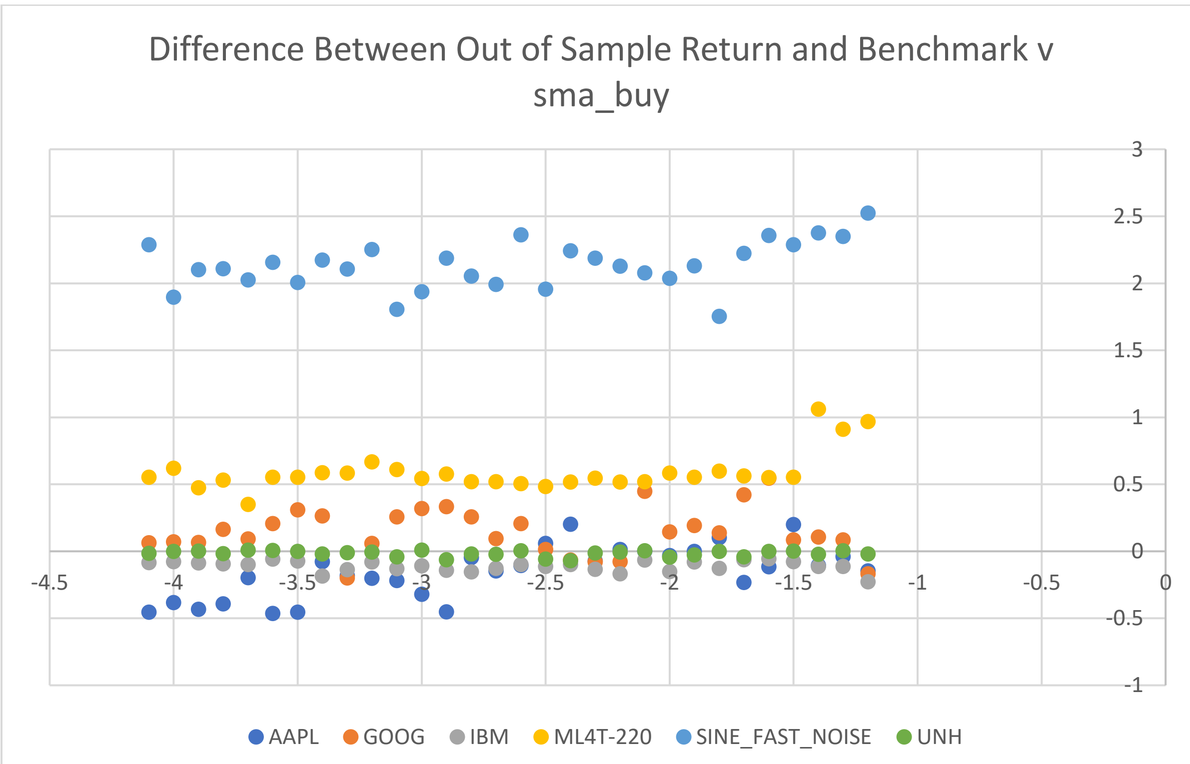
To, set up the range of features to feed to the learner, I first tried price to SMA ratio, Bollinger Band, and Momentum. Taking the rolling mean of stock prices, and dividing them by the current stock price, helped to make a prediction on whether to trade or not. The indicator, price to SMA ratio was more important than just the SMA, as price to SMA ratio would give a comparison between the current price and the average price in the recent past. Next, I used the feature Bollinger Bands. Theoretically, if there was a change in stock price greater than or less than 2 standard deviations from the SMA, then it would indicate a potential trading time. Next, I applied the momentum feature. I calculated the stock price ratio between the current stock price and the price before certain trading days. I then subtracted 1 from the ratio. I then normalized all three features and fed the normalized versions as X data to the learner. Once all three features were fed to learner, I calculated the Y data, which indicated whether to buy shares, sell shares, or to do nothing depending on some threshold values of the three indicators. Feeding all this data to the learner and then having it predict the trades for future data set was the plan.

Lastly, in the test policy function I looped through the trade on each day and if the cumulative holdings exceeded the max holdings of 200, I made an opposite trade of 200 shares on the previous day and continued on with the current trade. Another approach would have been to eliminate the current day trade entirely if the cumulative holdings exceeded the maximum holdings. However, I would lose out on some important trades that way and would lose out on potential profits. To keep up with the trades that the learner best predicted, I decided to use the former approach.

After some tweaking of the threshold values of each of the three indicators, I was able to pass all four tests on the grader with all out-of-sample cumulative returns positive and greater than the required returns, and all in-sample returns positive and greater than the required returns. It later turned out that adding one more feature volatility, which is the standard deviation of daily returns, helped me improve both the in-sample and out-of-sample cumulative returns and beat the required returns by greater margins.

Some of the hyperparameters that I tweaked were the threshold value of buying trades for SMA, the threshold value of selling trades for price to SMA ratio, same values for Bollinger Band and Momentum, the lookback period which was used to calculate the indicator values, the leaf size of the Random Tree learner, as well as the number of bags in the bag learner. I will only discuss 2 to 3 of the hyperparameters mentioned above. The three most important hyperparameters that I tweaked, were, the threshold value of price to SMA ratio for making a buy trade, the threshold value of the price to SMA ratio for making a sell trade, and the lookback period of the indicators.

The buy threshold value for price to SMA ratio could have been set to the minimum value of the feature for the given stock symbol. At all points where the feature value was equal to the minimum, a buy trade would be made or at points where it was less than a value close to the minimum a buy trade would be made. However, upon trying this inconsistent results were produced and it was hard to determine what percentage of minimum value would work for all symbols. So, I decided to set it to a fixed number. To find the target number I varied the price to SMA ratio buy threshold values from -1.2 to -4.1 with decrements of 0.1 for certain stock symbols and examined the cumulative return in-sample and out-of-sample. The following plot shows the difference between the out-of-sample cumulative return and the out of sample benchmark return versus the price to SMA ratio buy threshold values. The benchmark indicates buying 200 shares on the first trading day of the test period and selling 200 shares on the last trading day of the test period. Here, the other hyperparameter values were as follows: sma_sell threshold = 1.429, lookback period = 19. I picked these values, as they seemed to produce the desired results for me after some initial manual tuning.



The plot shows that the optimal value for the buy threshold occurs (value at which close to maximum out-of-sample return is achieved for most symbols) between -1.3 and -1.5.

I now incremented the price to SMA ratio sell threshold value to 2.429 and kept the lookback period value of 19 and ran the same experiment again. I did this with two more price to SMA ratio sell threshold values of 1.929 and 2.929 with keeping the lookback period to 19. Next, I also varied the lookback period with values of 13, 16, 19, and 21 and tested with price to SMA ratio sell threshold values of 1.429, 1.929, 2.429, and 2.929, with the price to SMA ratio buy threshold values ranging from -1.2 to -4.1 with steps of 0.1.

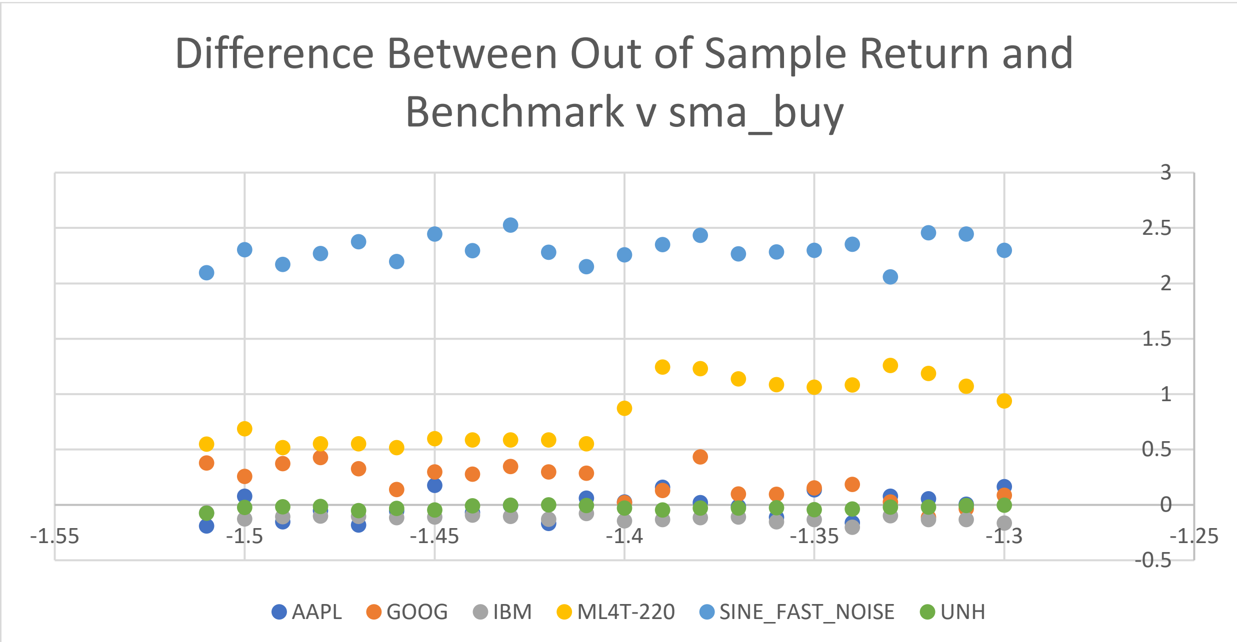
It turned out that almost all of them had the optimal price to sma ratio buy threshold value between -1.3 and -1.5. I have shown here the difference between out-of-sample return and out of sample benchmark for all stock symbols for these buy threshold values that happened to generate the best returns for all stock symbols. I can provide all the results upon request. The results are summarized below:

period	sma sell	sma buy	AAPL	GOOG	IBM	ML4T-220	SINE FAST NOISE	UNH
21	1.429	-1.5	-0.053	0.009	-0.108	0.732	2.216	-0.040
		-1.4	-0.092	0.259	-0.166	0.662	2.497	-0.039
		-1.3	-0.043	-0.061	-0.190	0.525	2.609	-0.031
	1.929	-1.5	0.089	0.334	0.054	0.772	2.227	-0.029
		-1.4	0.003	-0.030	-0.020	0.710	2.490	-0.064
		-1.3	-0.292	0.027	-0.111	0.594	2.560	-0.036
	2.429	**-1.5**	**0.114**	**0.405**	**-0.083**	**0.772**	**2.387**	**-0.046**
		-1.4	**0.117**	**0.374**	**-0.075**	**0.707**	**2.399**	**-0.057**
		-1.3	**-0.122**	**0.317**	**-0.032**	**0.564**	**2.662**	**-0.040**
	2.929	-1.5	-0.168	0.090	-0.059	0.772	2.015	-0.043
		-1.4	-0.241	0.304	-0.134	0.557	2.309	-0.023
		-1.3	-0.195	0.743	-0.092	0.564	2.341	-0.049
19	**1.429**	**-1.5**	**0.201**	**0.086**	**-0.076**	**0.554**	**2.288**	**0.000**
		-1.4	**-0.105**	**0.106**	**-0.114**	**1.061**	**2.375**	**-0.023**
		-1.3	**-0.038**	**0.085**	**-0.112**	**0.911**	**2.349**	**0.005**
	2.429	-1.5	0.024	0.493	0.015	0.626	2.211	-0.043
		-1.4	0.135	0.370	-0.037	0.626	2.293	-0.037
		-1.3	-0.207	-0.107	-0.034	0.730	2.094	-0.036
	1.929	-1.5	0.086	0.430	-0.016	0.662	2.284	-0.026
		-1.4	-0.100	0.572	0.054	0.662	2.102	-0.051
		-1.3	0.123	0.214	0.013	0.662	2.124	-0.025
	2.929	-1.5	0.208	0.365	-0.031	0.524	1.913	-0.054
		-1.4	-0.035	0.040	0.023	0.763	2.161	-0.040
		-1.3	-0.073	0.550	-0.006	0.662	1.876	-0.020

16	1.429	-1.5	-0.056	0.261	-0.139	0.507	2.336	-0.035
		-1.4	-0.224	0.186	-0.101	0.656	2.405	0.014
		-1.3	-0.347	0.147	-0.175	0.419	2.538	-0.040
	1.929	-1.5	-0.116	0.443	-0.068	0.819	2.038	-0.006
		-1.4	0.235	0.255	0.026	0.556	2.227	-0.028
		-1.3	0.058	0.494	0.003	0.374	2.236	-0.009
	2.429	-1.5	-0.041	0.299	-0.038	0.648	2.052	-0.011
		-1.4	-0.215	0.588	0.031	0.599	2.273	0.008
		-1.3	-0.004	0.135	-0.041	0.545	2.272	-0.037
	2.929	-1.5	0.069	0.344	-0.006	0.657	2.084	-0.012
		-1.4	-0.062	0.391	-0.029	0.599	2.140	-0.016
		-1.3	-0.211	0.532	-0.094	0.475	2.420	-0.038
13	1.429	-1.5	0.086	0.183	-0.145	0.569	2.436	0.010
		-1.4	0.047	0.037	0.003	0.556	2.388	0.004
		-1.3	0.095	0.248	-0.092	0.623	2.572	0.018
	1.929	-1.5	0.225	0.493	0.028	0.569	2.141	0.019
		-1.4	0.052	-0.077	0.040	0.500	2.184	0.013
		-1.3	0.119	0.244	-0.059	0.515	2.308	0.000
	2.429	-1.5	0.245	0.806	-0.090	0.720	2.060	-0.022
		-1.4	0.005	0.027	-0.038	0.500	2.453	0.012
		-1.3	0.100	0.198	-0.023	0.575	2.597	-0.003
	2.929	-1.5	0.190	0.816	-0.094	0.569	2.247	-0.026
		-1.4	0.078	0.201	-0.067	0.692	2.263	-0.026
		-1.3	0.159	0.591	-0.074	0.515	2.493	0.004

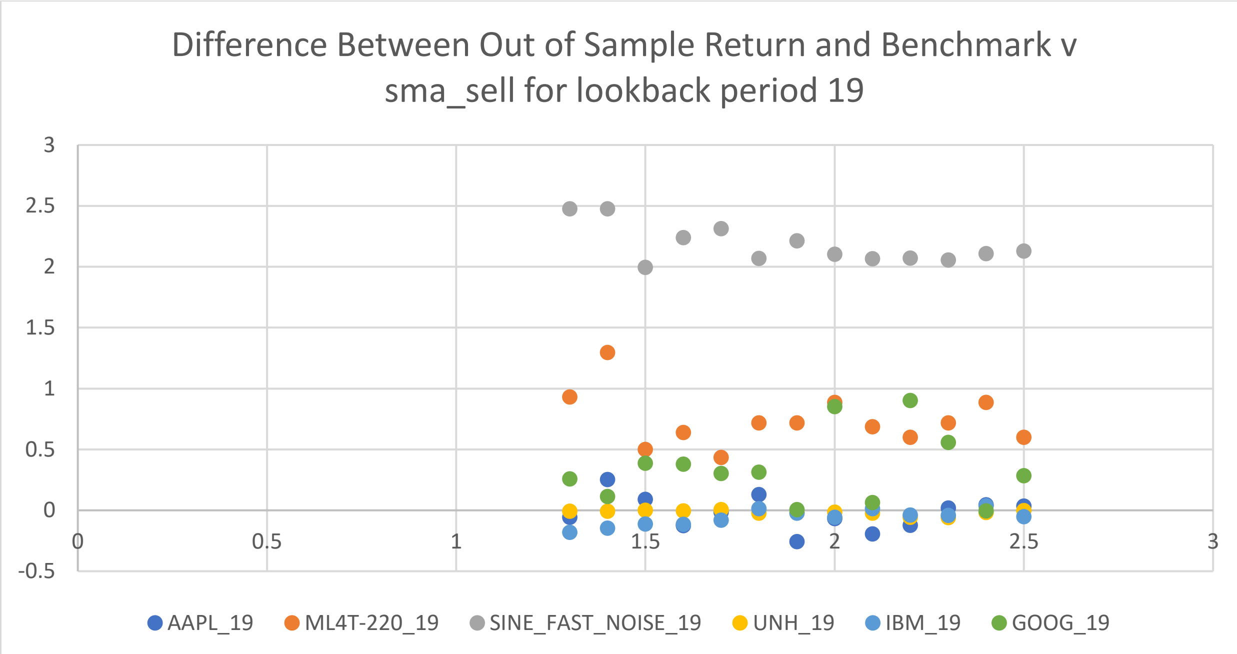
The table above shows that there were couple of sell threshold values that produced good results along with certain lookback periods. I have highlighted them in bold. The value 1.429 with period 19 however, seemed to produce better results that were in line with the four important stock symbols of AAPL, UNH, ML4T-220, and SINE_FAST_NOISE. Hence, I chose that as well as the lookback period value of 19 for the final set of buy threshold value experiment.

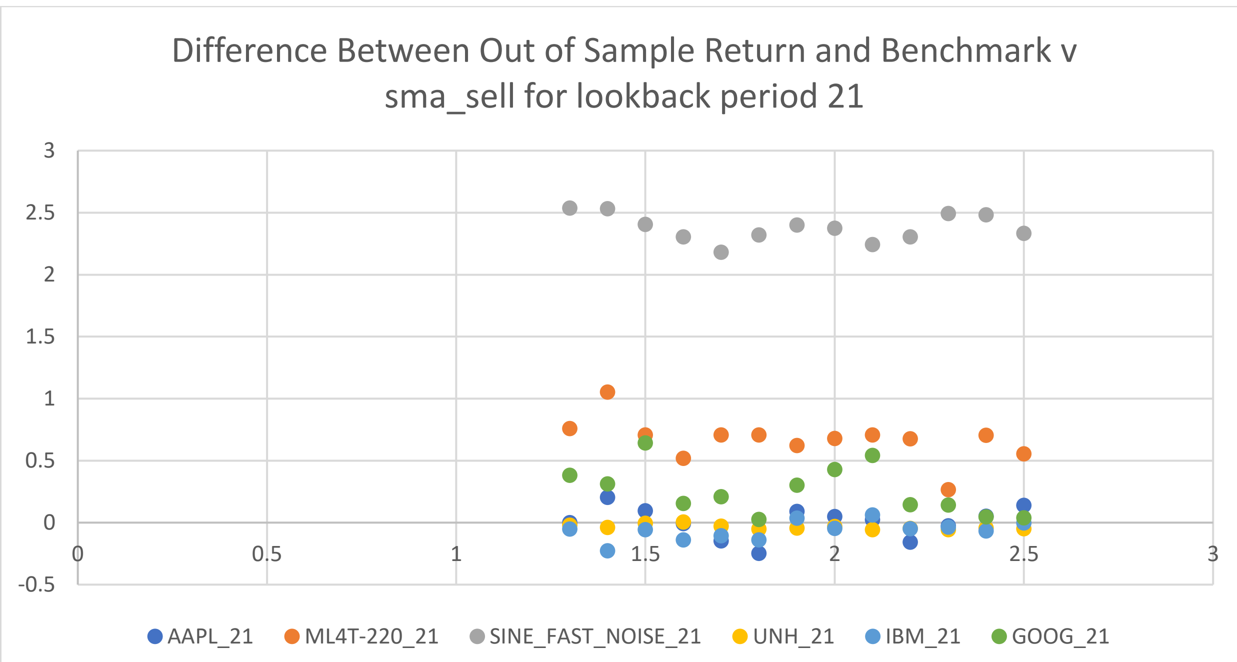
I ran the experiment with buy threshold values ranging from, -1.3 to -1.5, with decrements of 0.01. The results are shown by the plot below.



As the plot above shows, it turned out that the buy threshold value of -1.39 produced generally better results than other values. Hence, I settled with the value of -1.39 for the buy threshold.

The next hyperparameter that I worked with was the sell threshold value of the same feature. Here, using the same approach I would have liked to use the maximum value of the feature, and make sell trades during all those days. However, the number of trades would be greatly reduced. Hence, I decided to examine the price to SMA ratio sell threshold values for certain stock symbols and the resulting out-of-sample cumulative return. The above experiments showed that sell threshold values of 2.429 and 1.429 along with lookback period 21 and 19 respectively produced good results. So, I picked the range 1.3 to 2.5 for the sell threshold and periods of 19 and 21 with buy threshold of -1.39 and ran the experiment. The results are shown below.

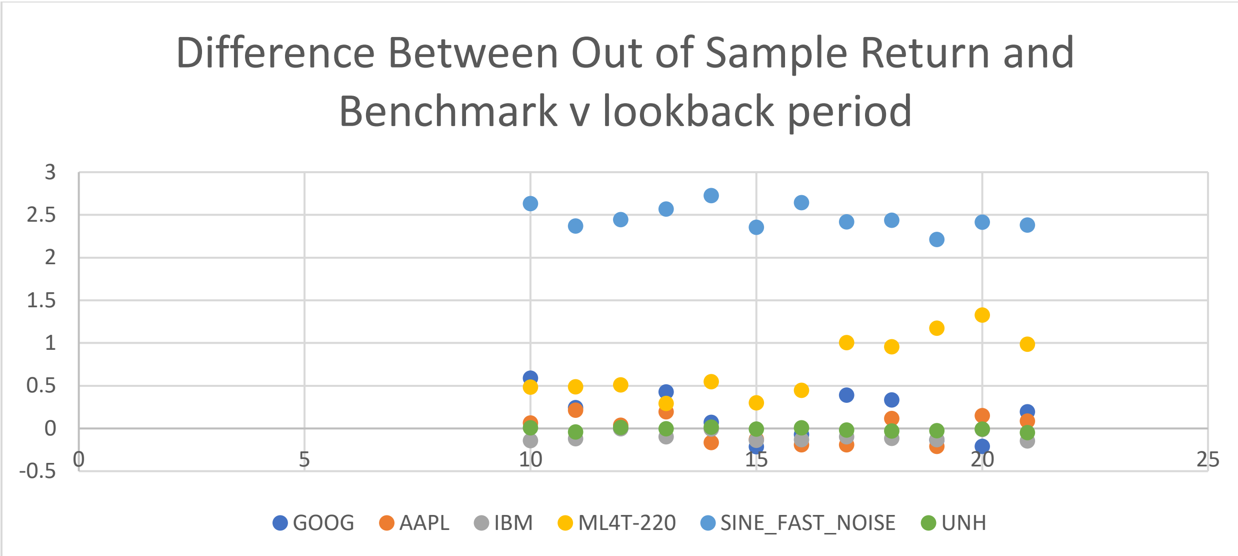




Both plots above show that the optimal sell threshold value occurs at around 1.4. For a more precise value, so that the grader would pass all tests, I tried the range of values 1.35 to 1.45 and then 1.425 to 1.435, and found that 1.429 was the value that produced the best results. I can provide the results if needed.

The next hyperparameter that I tuned was the lookback period for calculating the indicator values. I varied the lookback period value from 10 to 21 and examined the in-sample return and out-of-sample cumulative return for certain stock symbols versus the lookback period. The following plot shows the results with buy threshold value of -1.39, sell threshold value of 1.429.

Report



The plot above shows that values of 17 to 21 produced good results for ML4T-220, values of 11,13,18,20, and 21 produce good results for AAPL, values of 10,14,16 produced good results for SINE_FAST_NOISE, and for UNH the results were uniformly distributed among all periods. Of these values, I chose 19 as that produces good results for some and average or above average for others and is the one that gave me the best in-sample results for most stock symbols.

All hyperparameter values chosen were first used to test whether the in-sample cumulative return beat the benchmark return and then were tuned to produce better out of sample returns.