Yehuda Wexler

AIM 5005 Midterm

This is an analysis of a stock trading algorithm created for the NY State Pension Fund. The Fund is tasked with generating consistent and positive returns for its constituents, therefore there is an emphasis placed upon consistent positive returns, rather than impressive short term returns. Given this fact, I would not recommend utilizing either of the current models I will be presenting.

Using a month’s worth of QQQ ETF data, I ran an analysis to determine the best trading model. While there were some impressive returns, most notably a 3.51% monthly return vs. a Buy and Hold strategy return of .96%. The accuracy rate for each of the models never eclipsed 52%. Now, let’s dive into the models and their performance. I will note that the results increased significantly when tuning the hyperparameters for each model to an 80/20 train/test split.

I downloaded the data from Yfinance and have attached it to this submission.

Models:

* Random Forest
  + Simple (SMA, Standard Deviation, RSI)
  + EMA
  + Volume

I selected Random Forest because it works well with large numbers of features and non-linear relationships, as well as being quite flexible in the way it learns the data. I worked with n=100 samples since attempts with greater bootstrapping did not show any significantly better results and were much more computationally expensive.

* Logistic Regression
  + Simple
  + EMA
  + Volume

Logistic Regression is a classic, straightforward model, and is well suited for short term stock trading. I wanted to avoid getting into the Linear Regression rabbit hole and focus instead on classifying the data as up or down. The model was set up to select the best parameters for itself, such as the regularization strength (.01–10), penalty (Lasso vs Ridge), and liblinear as the solver to work through the calculations.

Interestingly, as you will see in the code, the model found different optimal regularization levels for each of the parameter tests. This model did not seem to work very well when Volume was introduced into the equation and the regularization spiked to 10.

Engineered Features include:

* SMA 5 Day
* SMA 15
* Rolling Standard Deviation 5
* RSI
* EMA 5
* EMA 15
* Volume Ratio 5
* Volume Ratio 15

SMA, EMA, RSI, and STD are all classically used to value stocks, which is why I selected those. Volume seemed interesting to throw into the loop (as it did turn out to be given the differences in performance between Random Forest and Logistic Regression)

I evaluated the models first by performance against the benchmark. As mentioned above, the best model achieved 3.51% return against the .96% buy and hold strategy. Coincidentally, this was the simple Random Forest model. Then, by ACcuracy, and finally, with a sanity check of the F1 scores.

|  | Accuracy | F1 0 | F1 1 | Return | Benchmark | Treynor Ratio |
| --- | --- | --- | --- | --- | --- | --- |
| Random Forest | 50.97% | .52 | .49 | 3.51% | .96% | 5% |
| EMA | 49.68% | .51 | .48 | 1.98% | .96% | 2.05% |
| Volume | 51.80% | .52 | .51 | 3.30% | .96% | 4.52% |
| Logistic Regression | 49.55% | .52 | .47 | 1.18% | .96% | .44% |
| EMA | 48.19% | .51 | .45 | .59% | .96% | NA |
| Volume | 50.71% | .43 | .56 | .98% | .96% | .04% |

As you can see in this table, the Logistic Regression model severely underperformed Random Forest in returns, F1 scores, and Accuracy.

Given these results, I would recommend further research into Random Forest, utilizing a larger dataset and experimenting with more features and hyperparameter tuning. As mentioned above, I do not believe this model is production ready, although the simple Random Forest model provided the best balance between consistency and returns– as evidenced by the Treynore Ratio ((Return-risk free)/risk). And this would warrant some more research.