Matthew Miller

# Final Report:

# Algorithmic Trading with Machine Learning

## The Problem

Intuition Technologies, a quantitative trading firm, manages assets exceeding $5 million as a private friends and family fund and would like to develop a sophisticated algorithmic trading strategy with machine learning models. The new strategy should be capable of identifying sizeable opportunities in the stock market and generating returns greater than those available from simply investing in the S&P 500.

In equity and financial markets across the world 60 to 75 percent of overall trading volume involves algorithmic trading. The percentage of algorithmic trading strategies have continued to rise in the 21st century as institutions appreciate the advantages of such an approach. Machine learning has shown promise in time series analysis and can potentially increase the profitability of an algorithmic trading strategy. As markets become more efficient it is essential to employ every tool available to a portfolio manager to maximize returns for the investors.

As with all investment opportunities the goal for Intuition Technologies is to maximize risk-adjusted return. As a data scientist, I have developed the foundation for a quantitative strategy that will meet the goal set out by Intuition. I have simplified the execution strategy based on the available information and predictions made by the machine learning model and will back test the performance of the model by showing the equity curves of investing in the portfolio generated by our model and that of the SPY index. The models will identify opportunities for return by predicting the direction of each stock the following day.

## The Data

I have compiled the end of day stock data for one hundred different anonymized companies. The data is stored in a different .csv file for every company, and each .csv file contains the date, open, close, high, low, and volume for each day from the beginning of 2010 until the end of 2019. In addition, we have the end of day data for the SPY from 2010 to 2019. Overall, the data is sized at around 11.0 MB. I have derived all technical indicators, performance metrics, and back testing prices from the close prices associated with each company's end of day data. I have left room for exploration as this data will focus only on unlevered long-only stocks to keep the calculations simple.

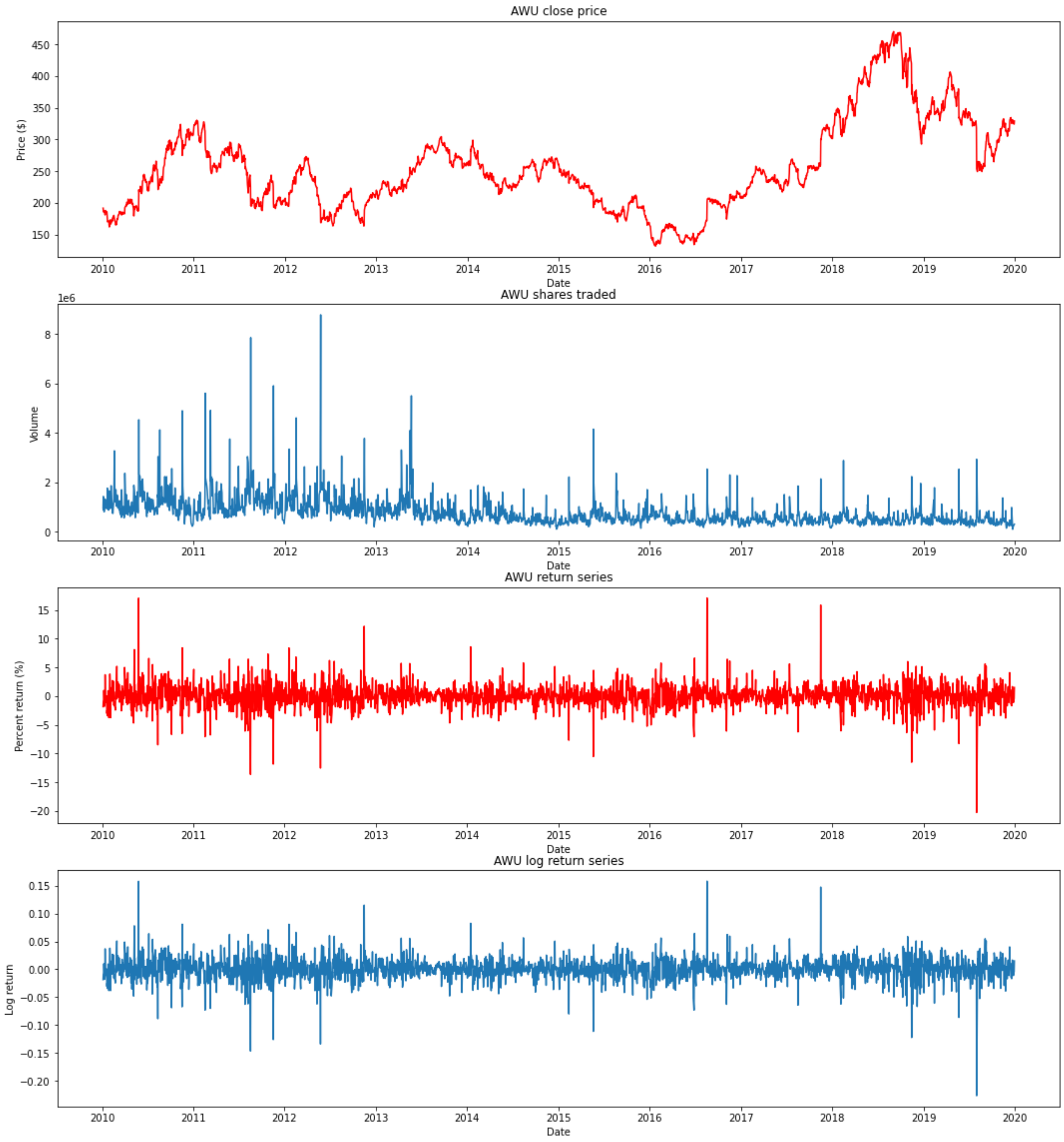


**Figure 1**. Single stock close price, volume, and percentage return series

The figure above displays the close price volume and percentage return series for a single stock of the 93 companies we have gathered and processed data for. The end of day close price and volume, or quantity of shares traded, were used to derive necessary performance metrics and technical indicators. The bottom two plots show the percentage return series and log percentage return series, which are series of price changes on the asset. Return series are great for deriving technical indicators and can be processed as signals unlike the price series. Although the percentage return series are more interpretable, the log return series are commonly used throughout algorithmic trading for their unique mathematical principles, which importantly includes symmetry of price changes.

## Exploring the Data

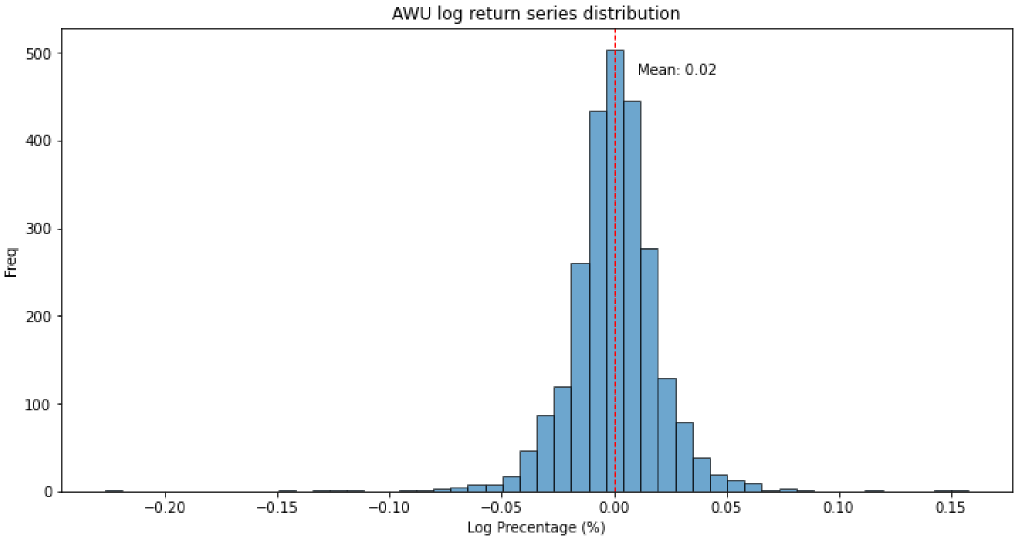


**Figure 2**. Single stock log return series distribution

First, I wanted to take a look at the distribution of the stocks to determine the type of distribution the series. Financial assets and their return distribution have been studied in detail for a long time. The prevailing theory for most financial assets is that they are log-normally distributed with fat tails. The signals derived from financial markets can be characterized by constant noise followed by occasional spikes and crashes, and therefore log returns with normal distributions are well adapted to highlight the extreme outcomes. In addition to the extreme outputs that can be seen in the distribution of the stock above, the average log return is around 0.02%, confirming the average price change for the stock over the period observed is positive.

The goal of this project is to maximize risk-adjusted return, and therefore I generated performance metrics capable of measuring it. Many performance metrics were considered, such as the Sortino and Calmat ratios, however the Sharpe ratio was chosen for its familiarity and simplicity. The Sharpe ratio is the most popular performance metric and measures risk-adjusted return of an equity curve. As risk is an analogue of volatility, the more volatile the price series of the stock is, the lower the Sharpe ratio. Sharpe ratios between 1 and 2 are considered good and ratios above 3 are considered excellent. Below in Figure 3 you can see the equity curves for three different companies from 2010 to 2020 and their resulting Sharpe ratio. While CUU displays the best return over that period from a standardized price, BMG would be considered a better performing stock considering it has less risk associated with its return series. I will continue to use the Sharpe ratio to compare the portfolios generated by using the machine learning models.

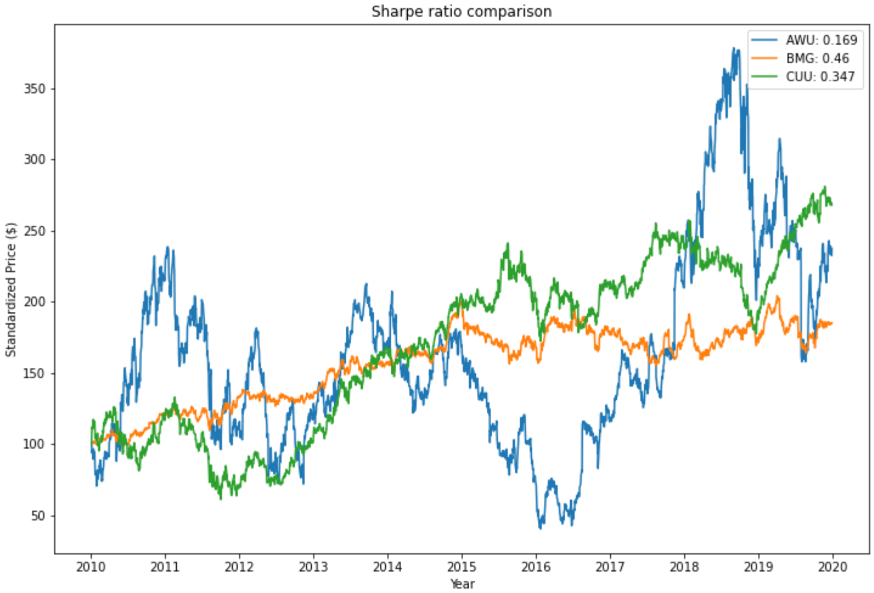


**Figure 3.** Sharpe Ratio comparison of 3 stocks

After I defined the performance metric to be used on our portfolios, I needed to determine the appropriate technical indicators or features for our models will train on. While performance metrics are used as objective variables for portfolio optimization, technical indicators are used as signals. These technical indicators are often used in algorithmic trading to discover patterns in market behavior and ultimately generate buy or sell signals. Creating these indicators requires moving averages generally created at 10 or 14 day intervals. However, for our pattern discovery we have used varying period lengths and allowed the machine learning models to determine the optimal period length through feature importance feedback. Below, Figure 4 displays three moving averages against the close price for a single stock. Note the smoothing factor characteristic of the larger windows applied.

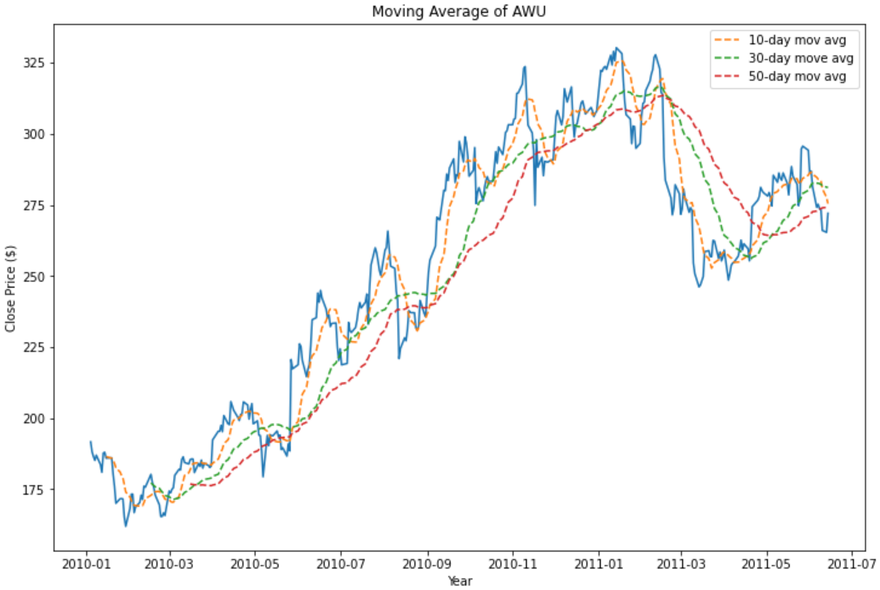


**Figure 4**. Moving averages at varying period lengths

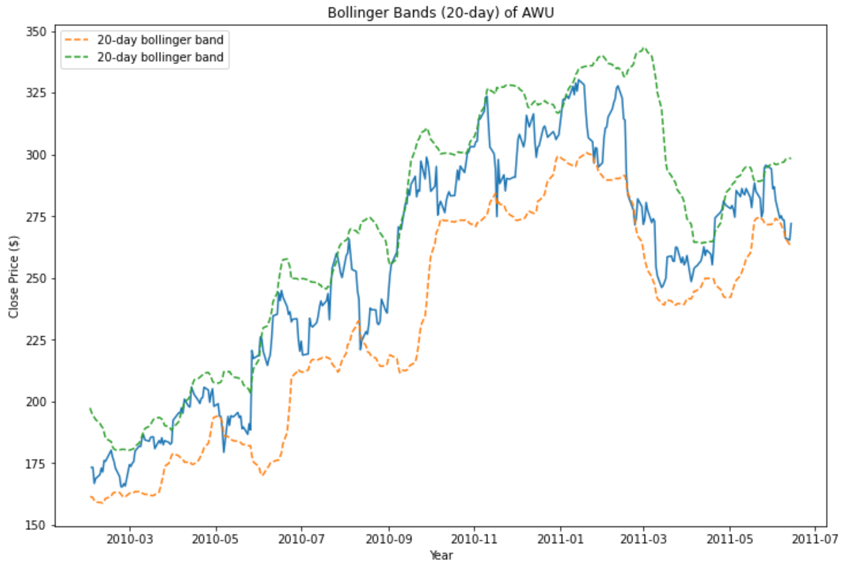


**Figure 5**. Bollinger bands with 20 days moving average

Above, Figure 5 shows the Bollinger Bands of a 20 day moving average, which are calculated as the second standard deviation over that period. Online brokerages often provide access to these types of plots, and traders will use indicators such as Bollinger Bands to make trades. The upper band indicates and "overbought" condition and provides a sell signal while the lower band indicates an "underbought" condition and provides a buy signal. For the purpose of this project I have created several more signals of varying period lengths, and allow the machine learning models to identify patterns for predicting the directional movement of the stock for the following day.

## Machine Learning and Results

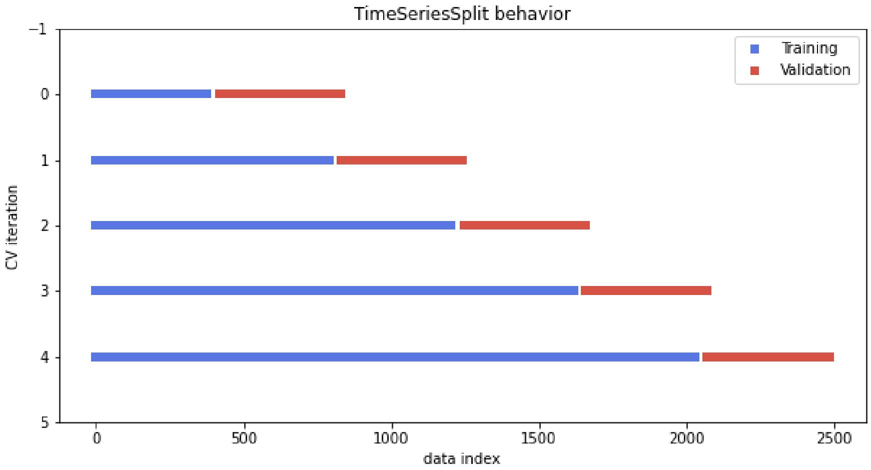
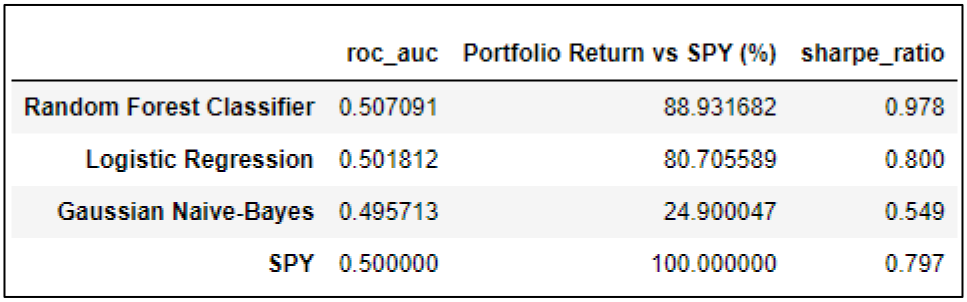


**Figure 6**. Time series cross-validation intervals

When splitting training and test sets of a time series dataset, it is important to maintain the sequential order. Therefore, while performing grid search cross-validation on our models, I used the TimeSeriesSplit from the sklearn.model_selection library. As you can see in Figure 6, using the TimeSeriesSplit iteratively progresses through a five-fold cross validation of the data. When applying the results of the grid search to our final model, I used an 80-20 split for the train and test sets. In addition, it was necessary to use an "embargo" technique to negate bias from the training data seeping into the test set. This was accomplished by dropping more observations in the data than the largest window size used in our rolling windows.



|  | roc_auc | Portfolio Return vs SPY (%) | sharpe_ratio |
|---|---|---|---|
| Random Forest Classifier | 0.507091 | 88.931682 | 0.978 |
| Logistic Regression | 0.501812 | 80.705589 | 0.800 |
| Gaussian Naive-Bayes | 0.495713 | 24.900047 | 0.549 |
| SPY | 0.500000 | 100.000000 | 0.797 |

**Figure 7**. Machine Learning performance metrics

Three machine learning models were tasked with predicting the directional movement of each stock's close price the following day. After tuning the hyperparameters and optimizing the feature selections based on the feature importance, the machine learning predictions were applied to the final two years of the data. The execution strategy was basic; if the stock was predicted to go up, then the simulation would purchase a predetermined number of shares. Above, Figure 7 shows the roc_auc performance for each of the models compared to a buy and hold strategy with the S&P500. Although the SPY proved the best investment compared to the machine learning models selected, the Random Forest which was the best performing of all the models, performed better when adjusting for risk.

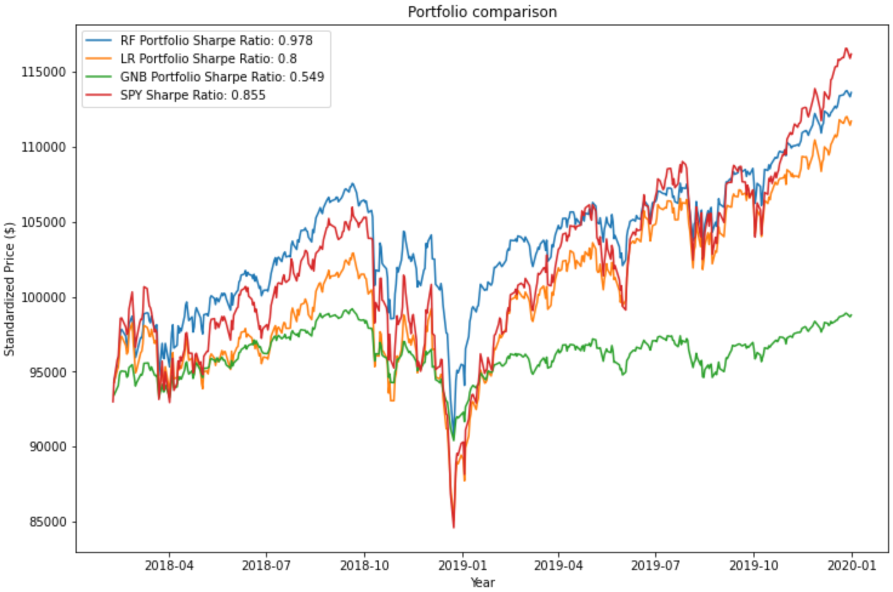


**Figure 8**. ML models vs SP500 equity curves

The above Figure 8 displays the equity curves for each model, showing the evolution of the portfolio over two years of predictions. As conveyed by Figure 7, the long hold strategy of the SP500 provided the best returns of all investment strategies with the Random Forest performing best of the machine learning models. However, with a Sharpe Ratio just below 1.0 the Random Forest model performed the best when adjusting for risk. The random forest and logistic regression models followed a similar shape as the SP500 while displaying less volatility.

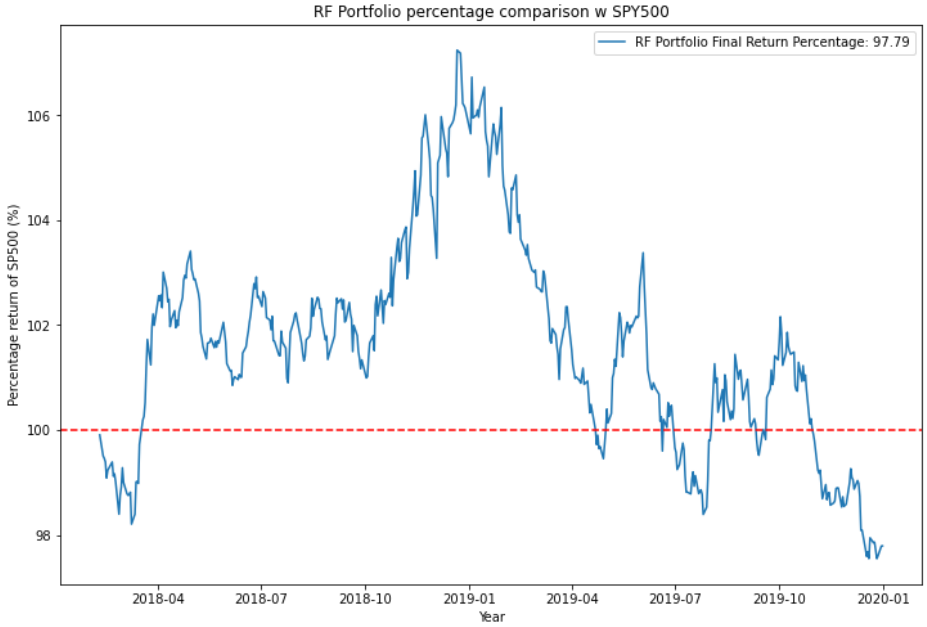


**Figure 9**. Random Forest return series as a percentage of SP500

A higher Sharpe ratio suggests a portfolio will perform more reliably over time. While the random forest failed to provide a better investment than the SPY return series, Figure 9 shows that for a majority of the two years back tested the random forest provided better returns. Therefore, after further consideration of trading fees and liabilities, I would suggest Intuition Technologies to continue the development of an algorithmic trading portfolio supplemented by machine learning. Machine learning clearly has potential in reducing the risk associated with investments, and potentially could be applied to ETF's or similar funds.

## Further Research

As mentioned, Intuition Technologies should continue researching the application of machine learning and data science to the financial markets. Some areas of particular interest would be levered assets, short trading, and the application of neural networks such as LSTM. LSTM models have shown promising performance on time series data and could improve our portfolio performance. In addition, further work could be done to improve the data collected. Collecting more data from further years past would improve the model's ability to identify patterns in the data, while exploring potential correlations between stocks within certain industries or market capitalizations could prove fruitful.