Applications of Machine Learning to Simple Financial Modelling

***Abstract:*** *Machine Learning is a powerful tool with many applications ranging from simple to complex purposes. The purpose of this paper is to show that simple analysis, with some key insight from machine learning, can result in useful models. I will apply machine learning algorithms to search for possible links between popular stock indices and I will build a financial model with the relations found as the foundation.*

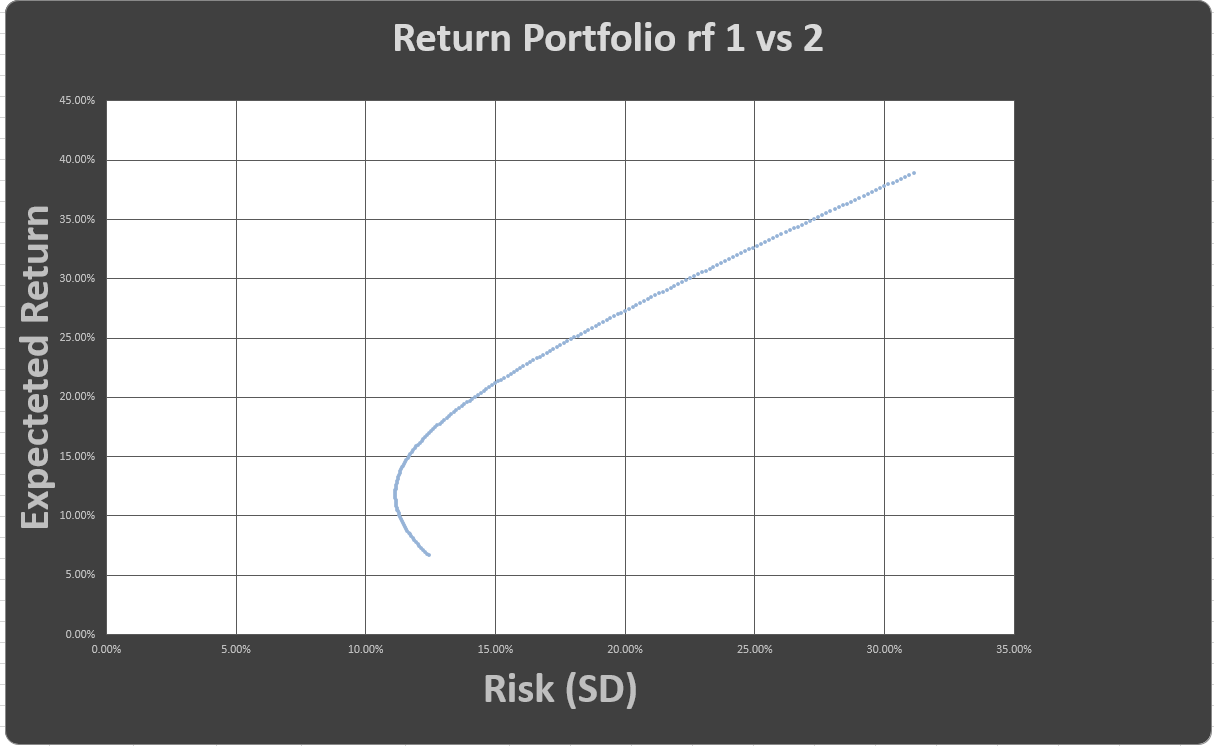


Figure 1 Example of financial model built from machine learning principle.

**Introduction:**

Machine learning is a very broad phrase. Applications can range from simple classification estimators to complex neural networks. Machine learning shows the potential to be incredibly powerful, however it is important to note the usefulness of the results. Does the new data tell you anything? What can be learned, or applied, from the conclusion?

It is important to keep these questions in mind while looking for machine learning to aid your work. The results need to tell you something new or useful, otherwise the machine learning process was pointless. Along with this, it is also necessary to use the appropriate estimator during the process, such that the most applicable and practical information is acquired.

I will attempt to find some sort of correlation between stocks, which is consistent and holds valid. Afterwards I will apply this new information by constructing a model which relies on any consistent correlation discovered from the machine learning estimator.

First I will analyze the magnitude and direction of change of the daily percentages of the parameter indices. This data will then be used to classify whether the target index moved upwards or downwards in price that corresponding day. The target data will be quantized to be binary, 1 if it moved upwards or change, or 0 if the stock moved downwards.

The stock indices were chosen very carefully and specifically. In order to assure that there is some sort of practicality of the results I gathered indices which are most commonly used to construct retirement accounts, such as a 401k or many Roth IRAs, and safe investment portfolios.

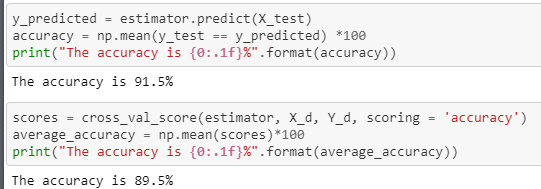
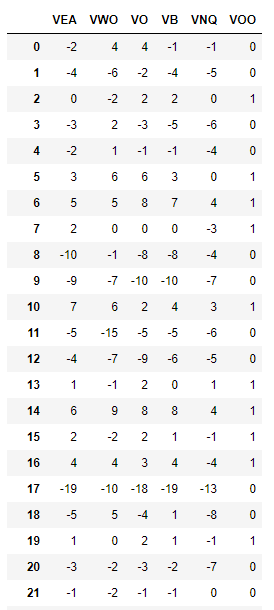
The parameter indices used are $VEA, $VWO, $VO, $VB, $VNQ, and $VV. The target index is $VOO. These are all Vanguard ETFs, which are very widely used for safe investment accounts for the general public. $VEA is the developed markets fund. This is a large mix of foreign company stocks such as Canadian, or European stocks. $VWO is the emerging markets index, which contains stocks from countries such as China, Brazil and South Africa. $VO is an index of ‘mid-cap’ US stocks.

**Focus:**

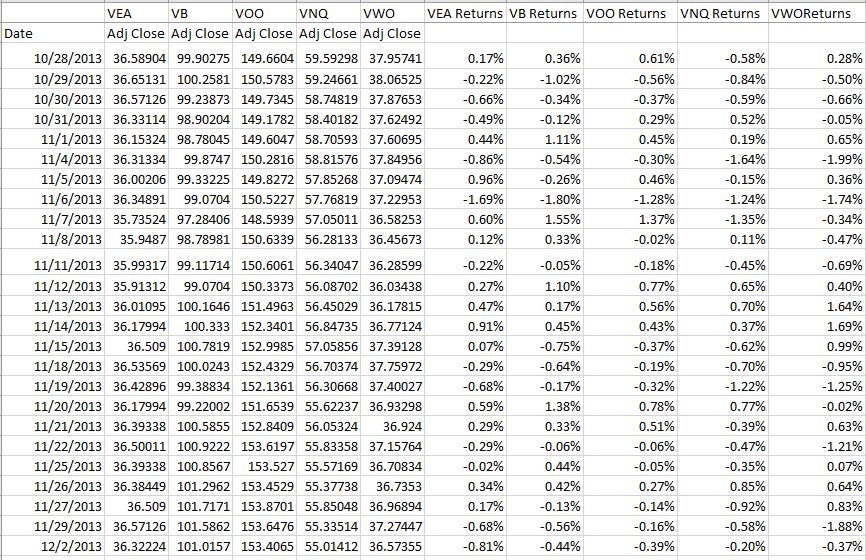
I will be analyzing approximately 6 years of daily historical data of 7 major stock indices. The daily historical data of each index will be then converted in to an array of daily percentage price change. 6 of the indices will be used as parameters for the machine learning estimator, whereas the direction of price change of the7th index will be used as the target data.

$VB is an index of ‘small-cap’ united states stocks, these are United States stocks which are within a certain market cap range. $VNQ is a real estate index, which is a blend of many large United States real estate stocks. $VV is the ‘large-cap’ stock index, which contains most of the largest publically traded United States stocks. The target index is an S&P 500 ETF, roughly, this is an average of the 500 largest United States publically traded companies. The S&P 500 is by many considered as the staple indicator of the overall stock market’s performance.

**Construction of Data:**

 As a csv file I downloaded the historical daily data of 6 years of trading, December 2012 to the current date, via Yahoo Finance. I then took the differences in in prices of consecutive days, in order to create a new column daily percentage changes in the respective index. Next, to make the machine learning process easier to apply I rounded the values of daily percentage changes to 0.1% accuracy. This way there is now a set of integers for the machine learning estimator use for training and testing.

As I wanted to find a correlation between the price changes of the given indices and the only the direction of change of the $VOO ( S&P 500 ) I converted the data of $VOO price changes to be either 1 or 0. The data value is 1 if $VOO movement was positive or 0 that day, and the data value is 0 if $VOO had downward movement. Now the data was ready for the estimator.

**Machine Learning:**

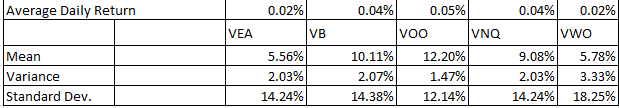
 I decided to treat this as a classification problem. I wanted the estimator to be able to distinguish only positive movement from negative movement, regardless of the exact magnitude of movement. The target data had only 2 classifications, either the movement is positive (1) or negative (0). The length of historical daily price changes of the parameter indices was 1510 data points. However, after splitting the data into training and testing data, via scikit-learn package, 1132 data points for training data and 378 data points for testing data. The estimator used was the KNeighborsClassifier from scikit-learn.

Table 1 Calculated values of Indices.

Table 2 Units being in 0.1%

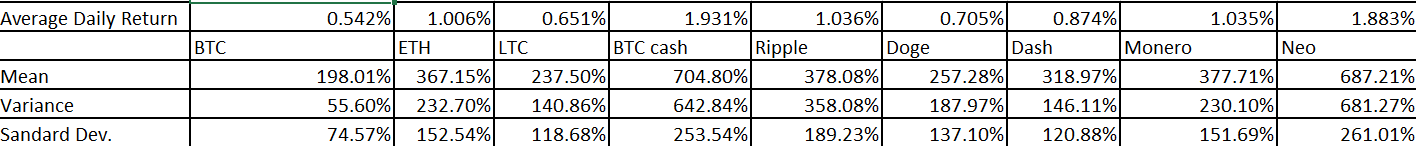
**Initial Results:**

Table 3Calculated values of Cryptocurrencies, notice inflated variance values and expected return values.

The accuracy of the estimator was 91.5%, with 89.5% accuracy with cross

validation, via cross\_val\_score in scikit-learn package. This implies a heavy correlation between the 6 indices and the S&P 500 ETF, $VOO. This means that the $VOO typically trends consistently with respect to the other 6 large indices. i.e. if $VEA trends upwards, $VNQ trends downwards, $VWO upwards etc, and $VOO trended upwards that day, this correlation tends to hold in the future.

If it is assumed to be true that there is a strong correlation between the indices, then a model which attempts to create a mixture of this set of stocks can somewhat predict how hedging can occur between these stocks inside of a portfolio as they move upwards or downwards.

**Hourly Data:**

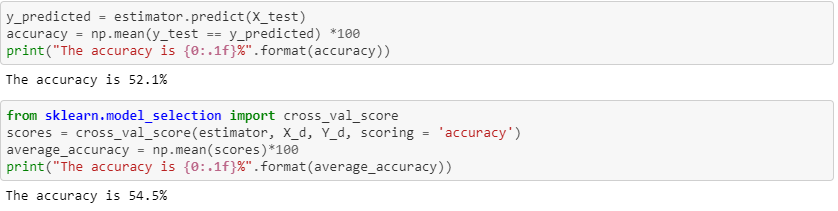
Noticing that the accuracy of the estimator’s results for the large stock indices on a daily scale are very high, ~90% I decide to start pushing the limit of the correlations. In order to have a stricter model, which could have better predictive power, I began to analyze the historical data on the hourly scale. This could have better predictive data, as this is a far shorter time scale for the price evolution of the stock, compared to the daily data. Less time for evolution of the prices into chaotic motion.

Unfortunately, there is no free downloadable data for hourly data for the stock indices used earlier. I decided to attempt the same process, however with more volatile equities. Cryptocurrencies had free csv data for hourly prices.

As this is now a shorter time scale, I decided to look for a correlation between the hourly prices of 5 different large cryptocurrencies vs. the price of Bitcoin the next hour. Before constructing the data sheets and applying the estimator it was understood that a lower time scale will result in less homogenous correlation between the cryptocoins vs that of daily movement, meaning that the predictive accuracy should be lower, however more useful. It is also noted that the cryptoc urrencies have a far greater price variance compared to the stock indices. This could result in a more difficult to predict model of short-term correlation.

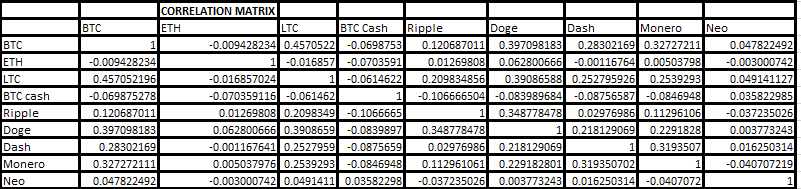
Cryptocurrencies tested include: BTC, ETH, LTC, BTC Cash, XRP, Doge, Dash, Monero, and Neo. BTC, Bitcoin, being the target data.

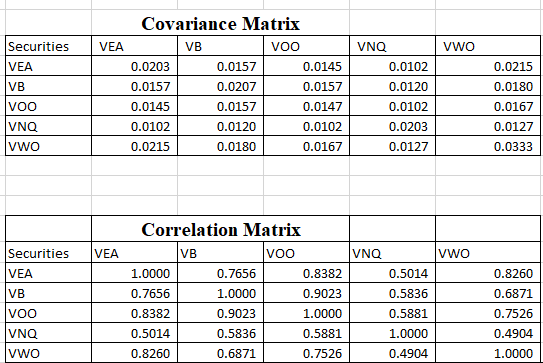
**Secondary Results:**



On the shorter time period the estimator had lower accuracy, as expected. The implication of the values for accuracy are great. The estimator had a greater than 50% accuracy for determining future movement of Bitcoin given the movement of multiple different cryptocurrencies of the past hour. This gives insight that some sort of correlation between the may be consistent. If it is taken to be valid that the correlation between these is valid then a financial model may be constructed on the basis that it could be assumed that correlation between the currencies is not random and can be treated as a valid, calculable variable in the model.

**Building the Financial Model:**

 Given myself enough evidence that correlation between the historical movement of these stocks and cryptocurrencies exist I proceeded to calculate the correlation, and covariance, matrix between all the equities in each of the respective machine learning results. It was also necessary to assign the success and risk of each respective equity by some metric.

 Using his historical expected return, and each variance of the equities as a metric of the risk and reward will give some sort of quantitative modelling of the quality investment portfolios crafted as a mixture of these indices and coins.

The usefulness of the information learned from the machine learning data comes in here. Since it is assumed that there is some calculable correlation between the indices and coins portfolios can be weighted based off of how the equities are expected to move with each other. Given that each equity has a unique variance, expected return, and correlation with all the other equities, the model has an extra dimension of depth. The model can now hedge the variance (risk), and the expected return (reward) by also including certain mixes of assets with strong positive upwards and downwards correlation.

Creating points along all possible, mixes of disjoint sets of assets the expected return of the portfolio and variance can be determined.

**Conclusion:**

Plotting the different portfolios results in a very interesting sideways parabolic graph. A key point to note is that there are certain portfolios which have the same the expected variance, however different expected returns. Implied is that, based off of the mathematical modelling and assumptions taken, there are objectively better portfolios to create with this analysis.

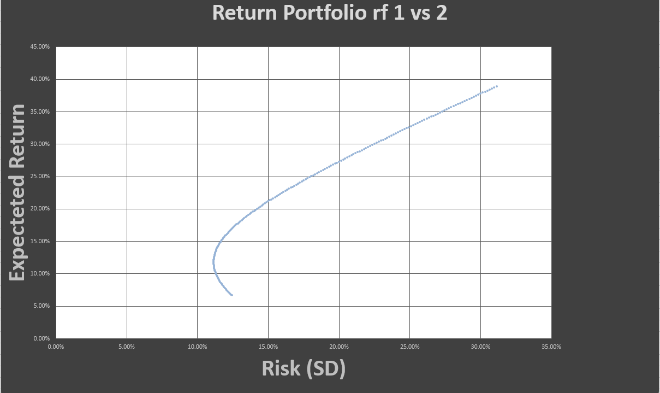
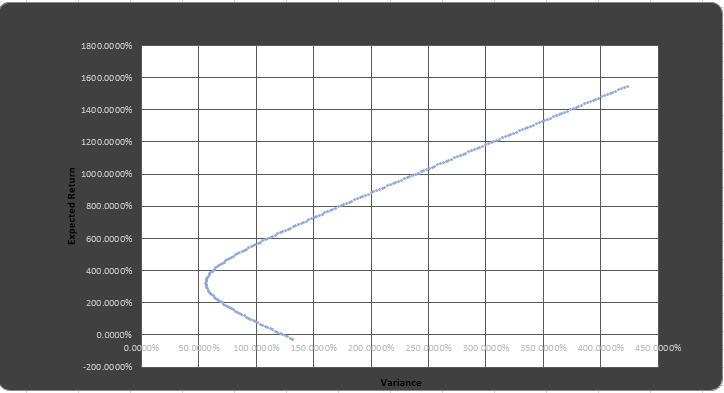


Table 4 Cryptocurrency porfolios

Table 5 Graphing all possible optimal portfolios of the indices.

It is important to note that the minimum for the variance of both the indices and the cryptocurrencies portfolios has a greater expected return than many other possible portfolios.

With the use of machine learning algorithms, I was able to find some correlations between assets. Given the strength of the machine learning estimator’s predictive power making well based assumptions of correlation I was able to construct simple financial models which provide useful information on comparing different investment account compositions.

All data was downloaded from

<https://coinmarketcap.com/>

and <https://finance.yahoo.com/>

for each respective asset listed.