International Journal of Mathematics Trends and Technology
ISSN: 2231-5373/ https://doi.org/10.14445/22315373/IJMTT-V69I9P502
Volume 69 Issue 9, 8-12, September 2023

Original Article

An Automated Statistical Learning Trading Model, Based on a Prior Data Study of the U.S. Markets from 1929 to 2019

Timothy A. Smith[1], Lauren Meyer[2]

[1]*Department of Mathematics, Embry Riddle Aeronautical University, Florida, United States.*
[2]*Graduating Student of MS Data Science, Embry Riddle Aeronautical University, Florida, United States*

[1]*Corresponding Author : smitht1@erau.edu*

Received: 13 July 2023 Revised: 19 August 2023 Accepted: 04 September 2023 Published: 22 September 2023

Abstract *- The phrase "one should not try to time the market" is some of the most valuable pieces of advice any new investor can hear, and while it is not desired to challenge that notion it is believed that portfolio rebalancing should be done when one sees danger on the horizon. In this paper, two main topics are discussed: Firstly, it is demonstrated through current data that the notion of 80/20 or 70/30 etc portfolio with the smaller portion being in bonds does carry risk to loss of principal in the bonds. Thus, an alternate investment instrument is proposed that yields the safety of a bond coupon, but also has potential for growth over time. Then, a mathematical model is presented to automate this method by the use of a statistical learning technique. The abstract theory of this method is presented in full detail, including a logical sketch of the algorithm utilized, but the exact algorithm is not presented as it is considered proprietary.*

Keywords *- Regression, Supervised machine-learning, Longitudinal, Macroeconomic, Trading model.*

1. Introduction

In this research study, a simple, yet highly effective portfolio management model, inspired by Buffet's advice "be fearful when others are greedy and be greedy when others are fearful," is presented using a formal mathematical procedure called statistical learning. Over the years a great amount of study has gone into constructing portfolio management techniques to optimize returns while mitigating risk; however, over recent years new information has come to show that the majority of long term gains have come from just a few companies, and in this study it is investigated how this new information can be applied. Moreover, in recent research [1-3] it was discovered that over the last century nearly 90% of the wealth in the American markets was created from 5 large companies, namely: Apple (AAPL), Amazon (AMZN), Google (GOOG), Microsoft (MSFT) and Exxon (XOM). Moreover, following recent data [4] these trends have continued into 2023 with a slight rotation in the leadership: Apple (AAPL), Amazon (AMZN), Google (GOOG), Microsoft (MSFT) and Nvidia (NVDA) dominating. Namely, at the end of Q2 Apple was holding $2.79T, with Microsoft holding $2.46T, Google holding $1.95, Amazon holding $1.25T and Nvidia holding $0.99T. If this is compared to the total estimated market cap of the SP 500, which was approximately $35T at the end of Q2, it shows that these five companies hold approximately 27% of the total market cap of the index; however, when investigating the actual growth in these companies over the time period of Q1 and Q2 of the current year, it is amazing to see that these five companies account for 96% of the SP 500's gains [4]. It goes without saying much that the investor who invested in these companies did very well, and would be expected to continue to do well in the future, but of course there is risk investing in any stocks, even these most profitable ones. Thus, in this study a mathematical model to optimize returns while balancing safety is proposed.

In this study, to begin it is illustrated that while these companies have strongly overperformed over the last decade, they do still carry risk which can be somewhat offset by including safer instruments like bonds into the portfolio. In addition, it is demonstrated through data of recent events that the generally accepted safe investment of bond funds may not be the most effective. For example, during the last few years the market has encountered extremely volatility and if an investor put $500,000 into just the five stocks (APPL,AMZN,GOOG,MSFT & XOM), "the big 5" henceforth, at the beginning of 2021 their return would have increased to $623,528 as of December 2022. On the other hand if a portfolio of 50% into the "big 5" and 50% into the safe assets, then the return over the same timeline would be a gain of $592,914. While of course the gain of

the big 5 is superior, it is not to say that there wasn't a gain on the safe assets; moreover, the question comes to mind that was it worth the risk to obtain that extra $30,614, or 3%, yearly, gain? Similar results can be seen over longer time period looking back at the same calculation before the pandemic investing in Jan 2019 the values would be $826,457 and $738,886 respectfully, or $730,502 and $646,120 if the investment was initiated in Jan 2018. Moreover, the notion of investing 50% (or whatever other percentage) into the bonds fund for safety may not always be as safe as one believes. For example, if at the end of 2022 the amount of $250,000 was invested into a common bond index held in many 401K or 403B type accounts, such as VBTLX or BND, as of January 2023 this amount would have declined to $193,565 or $205,073 respectively, which is an average loss of 20%. Obviously this is an extreme valuation which is a somewhat unforeseen side effect of the current raising rate environment, but it happened and it is quite possible that future events could cause similar phenomena.

In this research two financial instruments are investigated: Firstly, the growth assets here is a simple equal weighted index referred to as "the big five less one," which is just an equal weighted portfolio of four stocks; moreover, while here the raw data of just four stocks is used, in most applications this could be replaced by commonly used broad market mutual funds. Moreover, it was specifically chosen to not take all five of the stocks as it is expected that as time moves forward companies will change with some dropping from leadership while others moving up into leadership, but by using common sense one should be able to identify the ones that are still the "core of the core." Then, an alternate method to bonds is suggested which is a subset of the Dow Jones, but only including companies that have maintained, or raised, their dividends in the last year. Within this paper, publicly available data for the exchange traded fund from the Invesco Dow Jones Industrial Average Dividend ETF (DJD) are used to represent the second asset. Hence, the DJD instrument is replacing the bond funds commonly used, with its dividend being an amount comparable to a bond coupon.

The goal of this portfolio management strategy is hold the majority of the portfolio, if not 100%, in the "big five" while times are good and the market is steadily growing, but when the market begins to fall investors should then move to safety through the "bond type" asset. A counter argument to it being safety is that fundamentally it is stocks, and while it is true that it is stocks there is more to the story than just that; namely, the DJD fund pays a dividend so it guarantees a cash flow, currently that value is approximately 3.5% yearly. In addition, the stocks contained in this instrument are very large stable companies, basically the United States economy, so the likelihood of it loosing value in the long run is near zero! The data illustrated that while it does move up and down with the economy, generally the magnitude of the moves is smoothed out. This is observed with the instrument not falling quite as sharply during the market panics of the pandemic during March 2020 or the recent war/oil fear of October 2022.

The most important concept here, is while it is true that the "bond type" asset does carry risk to decline in the short term, it is almost certain to continue up to the right in the long run; hence, this asset gives the guarantee of 3.5% yearly yield, similar to a bond, but also the assurance of long term growth of principals. This should be very attractive to an investor at any time, but especially at current time as many bond funds loosing value!

1.1. Data Introduction

Annual rebalancing of a client's portfolio is very important for long term wealth growth, and this research illustrates the importance of flexibility in the portfolio to adapt to the market when it adjusts and changes. Generally most current financial advisors will suggest that the most effective long term trading strategy would be an all-in investment into the "big five" stocks, or just an overall market instrument such as an SP500 index fund, while the investor is of younger working age. Then transition the portfolio slowly to an 80/20 then 70/30 etc, as retirement grows closer. However, upon completing the data investigation for this project this was found not to be the case. For example, if a portfolio of $500,000 in the "big five" was studied from the initial investment time of January 2021 to December 2022, the value at the end of 2022 would be $623,528. Likewise, during a longer period, from January 2019 to December 2022, its value would be $1,029,461. Although this still seems like an obvious choice to stay "all in," there must still be a concern for mitigating the risk of these fluctuating investments in the long term. Clearly the time frame used is an illustration of when the market was performing well. It is prudent to consider a portfolio that must also include some other safe instrument. A commonly used bond instrument is the total market bond fund, BND, but if $500,000 was put into that instrument during the longer time period, it would have actually decreased to $448,218.

For further study in this data overview, "the big five less one" technique was attempted with a 50%/50% investment strategy. In 2019 to 2022, a new portfolio was created with 50% ($250,000) in the big five stocks and 50% ($250,000) in the more conservative "bond type" investment. From this 50%/50% strategy, the return was $826,457. This same method done in a 2021 to 2022 had a return of $592,913, this only was $30,614.26 less than a full investment in the big five approach discussed previously, but without the heavy risk. For comparison, if during the 2019 to 2022 time frame if the 50%/50% strategy was applied but using the BND instrument the value would be $738,840, approximately $87,000 less than the newly proposed

method. In summary, the newly proposed method is proving to be superior, and the icing on the cake would be to also identify a mathematical method to identify moments in time to reduce exposure to the risky assets as doing so would prevent losses, hence truly optimize returns!

2. Summary of work

Monthly data was collected [5] for a large time period covering almost three decades, going back to 1991, to create a regression model with the response variable here being the overall stock market. The data of the S&P 500 value was used for simplification, as it is generally accepted as a measure of the overall market and has data that is readily available. The predictor variables used were based on prior research ideas to model various macroeconomic indicators. In theory, any data could be utilized, but from prior experience [6-8], it is understood that a model with around 3 to 5 predictors should suffice; hence, to being data was collected [9-15] from government websites of Consumer & Producer Price Index, used to model inflation, and Gross Domestic Produce & Money Aggregate, used to model money into the economy, Fed Funds Rate, used to model borrowing cost, and Unemployment Rate, used to model workforce participation. Now, in practice, a common method would be just to dump all of this data into the software, and in doing so, a lasso method from Python yielded the model.

$$\hat{y} = 32.2913FFR + 0.22776MS - 98.89U - 3.2185PPI - 0.0166GDP$$

However, as expected, several of these variables have inter-correlation. Hence, prior to running the model using a traditional backwards subset regression, it was desired first to do a pre-correlation analysis, looking for any pair of predictor variables that had a cross-correlation $r_{i,j}$> 0.9 and removing the one from the pair with the highest variance inflation factor if it was greater than 10. Now, doing resulted in a four-variable model

$$\hat{y} = 613.6524 + 31.81978FFR - 0.02886GDP + 0.216199MS - 107.806U$$

which interestingly performed about the same as the prior model, both having coefficients of determination of approximately 97%. Therefore, it was preferred to use the latter four variable model as it is both more efficient and possibly removes multicolinearity issues.

Looking over the last two decades, the computations of this model performed very well! While here it is not attempted to look for "overvalued" market times, it is worthmentioning that prior to the "dot com crash" of 2000, this model was flashing that the market was overvalued by over 20%; hence, it predicted that it would crash back through mean reversal. Our point here is to identify buyback in times as this model should tell us if things will continue to trend down, think the predictors moving with, or if the bottom is near, think the predictors moving in other direction.

The following results were obtained in regard to our model's predictive "move back in" times and the market's actual "bottoms." For illustration, three recent major market events are considered: the 2000 dot com crash & 2008 housing crash, and the 2020 covid mini-crash, and a critical value is defined to be 10% below (which could be raised)

- In January of 2003, this model first showed that the approximated "y^ or y hat" values had crossed past 10% below when the market was around $900
  - and the actual market bottom was a few months later at a value of around $830
- In October of 2008, this model first showed that the approximated "y^ or y hat" values had crossed past 10% below, when the market was around $1000 and if
  - and the actual market bottom was a few months later at a value of around $800
- In March 2020, this model did not go past the 10% level, as it is only running once a month, on March 1st prior to the crash and April 1st, which was a few days after the market had two back-to-back daily gains of over 9%

3. Generalizations of Mathematical Model & Conclusions

In this section, a model is created to identify these peaks and troughs of the market, and of course, while it is not possible to exactly time the market, this model does work very well when it is applied for longer time frames. A linear regression model is created in the format:

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots$$

Here, the predictor values, x1,x2, x3, etc, represent data from macroeconomic predictors such as the normalized values of the Gross Domestic Product, the Consumer Price Index, the Fed Funds Rate, and other factors. In theory, any number of predictor variables of any data can be utilized, and a procedure is outlined to find the best model. It is expected that other

predictors, such as average national gas prices, might be desirable, too. However, caution is provided to the reader regarding overfitting a model by adding in variables that do not significantly increase the model's performance. Moreover, the equations here are outlined to show how this process works for linear regression, but other kinds of models could be created by utilizing the same logical procedure.

To begin, what is called the first level model is created, where only one predictor is entered into the model separately on its own but running one by one through all of the predictors. This will lead to a set of single variable models of the form:

$$y = \beta_1 x_1 + \beta_0$$

Here, β is the slope fitted from the software, and β_0 is the intercept, which is essentially the starting value of the y value. From all of the models created, the best model is defined as the one with the highest R squared, and the associated predictor variable is called x1.

Then, the second-level models are created, where two variables are entered into the model in a pattern. This is done by putting in x1 and pairing it with one other predictor entered into the model separately but running one by one through all of the remaining predictors. This will lead to a set of two variable models of the form.

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_0$$

Again, from all of the models created, the best model is defined as the one with the highest R squared, and the new associated predictor variable is referred to as x2.

The next step is to continue to add additional predictors to create more multivariable models following this pattern. This will lead to $n \cdot (n \cdot 1)$ two variables models, $n \cdot (n \cdot 1) \cdot (n \cdot 2)$ three variables models, and so forth. This process is continued for third- and then fourth-level models and continued until the data set is exhausted. For each level, the best model is selected – here, the R squared is used – and then a final model selection is made by analyzing the four best models. There may be some debate about how this final model selection is done. In most cases, a simple statistical measure can be used, and even with the growth of "big data", it is still preferred to use models with the minimum number of predictors; hence, one may construct a code that terminates when the delta in R squared is less than 0.01.

In summary, the methods in the prior sections provide either a model to be used at the current time or a procedure to create a model in the future, and this model can be used to identify points in time when adjustments to the portfolio are needed to/from safety to risk on. The desired portfolio would contain two elements: the first being the safety "bond type" asset, and the second being the "risk on" assets of the "big 5", both as previously described. The method of investing is to hold the majority of the portfolio in the second asset when times are well and then use the mathematical models described, in addition to common sense, to move to safety when the economy is either overheard (which the maths will identify) or when geopolitical risks cause concern. Then, the mathematical models can identify times to move back in. Nobody can, or should, perfectly time the markets, but for those who believe in mean reversal, this mathematical method will provide a solid procedure to follow for the long run. In addition, the proposed "bond type" asset has the characteristic of acting as a safe asset, with a yield similar to that of a bond coupon, but it will not miss out on all market gains; hence, in a sense, it can be viewed as providing both safety while lowering the risk of lost opportunity cost.

Funding Statement

Any grant-type mechanism did not fund the research leading to this article, but it was done as an MA 680 capstone research project course as part of the student's M.S. Data Science program.

References

[1] Hendrik Bessembinder, "Wealth Creation in the US Public Stock Markets 1926-2019," *The Journal of Investing*, vol 30, no 3, pp. 47-61, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] Atreyee Sinha Chakraborty, "Innovation, Technology Transfer and North-South Trade," *SSRG International Journal of Economics and Management Studies*, vol. 3, no. 6, pp. 11-15, 2016. [CrossRef] [Publisher Link]

[3] Ibuomo R. Tebepah, "Digital Signal Processing for Predicting Stock Prices Using IBM Cloud Watson Studio," *SSRG International Journal of Computer Science and Engineering*, vol. 7, no. 1, pp. 7-11, 2020. [CrossRef] [Publisher Link]

[4] SHAWN TULLY, Just 5 stocks including Nvidia account for 96% of the SP 500's gains this year, Fortune, 2023. [Online]. Available: https://fortune.com/2023/05/30/nvidia-spx-index-stock-market-outlook/

[5] S&P 500 Market Cap, 2023. [Online]. Available: https://ycharts.com/indicators/sp_500_market_cap

[6] Faisal Khan, "Macroeconomic Factors and Stock Market Inefficiency: Role of Trading Effect," *SSRG International Journal of Economics and Management Studies*, vol. 7, no. 3, pp. 8-14, 2020.[CrossRef] [Publisher Link]

[7] Rohit B R et al., "Stock Market Prediction using Machine Learning," *SSRG International Journal of Communication and Media Science*, vol. 7, no. 2, pp. 6-9, 2020. [CrossRef] [Publisher Link]

[8] Timothy A. Smith, and Alcuin Rajan, "A Regression Model to Predict Stock Market Mega Movements and/or Volatility using both Macroeconomic Indicators & Fed Bank Variables," *International Journal of Mathematical Trends and Technology*, vol 49, no 3, pp. 165-167, 2017. [CrossRef] [Publisher Link]

[9] Sanjoy Kumar Saha, Nilufar Easmin, and Partho Sarathi Laskar, "Gravity Model and Trade Flow of Selected 20 Countries," *SSRG International Journal of Economics and Management Studies*, vol. 5, no. 3, pp. 22-30, 2018. [CrossRef] [Publisher Link]

[10] The United States Consumer & Producer Price Index and Unemployment Rate, Bureau of Labor statistics website, 2023. [Online]. Available: http://data.bls.gov/

[11] The United States Gross Domestic Product, Bureau of Economic Analysis website, 2023. [Online]. Available: https://www.bea.gov/data/gdp/gross-domestic-product

[12] Andalan Tri Ratnawati, "The Effect of Trading Volume Activities, Earning Per Share and Stock Returns on Market Value Added (Empirical Study on LQ45 Companies)," *SSRG International Journal of Economics and Management Studies*, vol. 8, no. 2, pp. 79-82, 2021. [CrossRef] [Publisher Link]

[13] Thi Hoang Oanh Nguyen, "Preferential Trade Agreements: Development Trends and their Effects," *SSRG International Journal of Economics and Management Studies*, vol. 9, no. 4, pp. 1-16, 2022. [CrossRef] [Publisher Link]

[14] The United States Money Measures, Board of Governors of the Federal System website, 2023. [Online]. Available: https://www.federalreserve.gov/releases/h6/current/default.htm

[15] The United States Federal Funds Rate, H.15 Selected Interest Rates, Board of Governors of the Federal System website, 2023. [Online]. Available: https://www.federalreserve.gov/releases/h15/