Using Data Science to Build a Better Retirement Portfolio

Steve Donahue

In December 2007, a global financial crisis shook the world economy. A bubble in the American housing market broke, and triggered a chain reaction through the securities sector, which had many funds that were backed by the defaulting mortgages. The resulting ripple effects were widespread: jobs and homes were lost and the retirement dreams of many baby boomers were shattered. It was two long years before parts of the American economy began to recover, and even ten years later there are sectors of the American population who are still waiting for relief.

Americans have become too dependent on financial advisors and fund managers to shape their understanding of how to build wealth. These same financial professionals are incentivized to drive returns ever higher, which in turn places the money entrusted to them at greater risk. The common investor should be able to use the power of existing data to develop a practical and effective portfolio, which gains with the growth of the economy but does not collapse under a breaking bubble or recession.

This data science project seeks to replace conventional wisdom with data driven insights, creating portfolios from random stock indices, intermediate term bond funds, long term bond funds, gold-based funds, and commodity funds. The success of these portfolios is then evaluated with historical data to develop meaningful principles to mitigate risk while optimizing returns.

Methodology

The categories chosen for study in this project (stock indices, intermediate term bond funds, long term bond funds, gold funds, and broad basket commodity funds) were selected based on the All Weather Portfolio, a featured portfolio in Tony Robbin's Money: Mastering the Game. Money distributions and rebalancing schemes will begin with select options from the book and then hypotheses for improved performance will be tested.

Fund data will be collected from Yahoo: Finance (https://finance.yahoo.com). This site was selected because the data available is clean and comprehensive. Jupyter notebooks (and the Python programming language) will be utilized for the analysis.

To simulate a portfolio

1) Select one random fund from each category to form the portfolio.
2) Accept an initial investment amount and a recurring bi-weekly addition.
3) Accept a user specified proportion for the classes of the portfolio.
4) Evaluate the performance of the portfolio across historic data.
5) Rebalance the portfolio at a regular, user specified interval.
5) Output Compound Annual Growth Rate (CAGR)
6) Output the worst loss across 52 consecutive weeks.

To perform the data exploration

1) Create distributions of portfolios with various input parameters
2) Create distributions of portfolios which change under a variety of rules.
3) Develop hypotheses for improving on the All Weather Portfolio
4) Test the hypotheses against historical data

Statistical Methods

1) Simulated distribution will be tested for a mean of 9.7% CAGR
2) Simulated distribution of losses will be tested for a max loss of 4.3%

Lastly, the code which supports this analysis may be used further to create an interactive tool that any curious investor could easily use to learn about broad fund investing, test parameters of their own choosing, and gain insight into the behavior of the economy. Predictive classification tools may also be trained on the existing data using time series methods to determine likely future prices for the different categories. If effective, these classifiers will also be used to build a response mechanism within the model to move money between funds at advantageous times.

The Data and Simulated Portfolios

The abbreviated following funds were identified within each category by the researcher following a review of available investment publications.

Stock Indices: FUSEX, PREIX, SWPPX, VFINX, VIGRX
Intermediate Term Bonds: BIV, HYG, IEF, IEI, IGIB, IPE, ITE, TIP
Long Term Bonds: PRULX, VUSTX, WHOSX
Gold: INIVX, OPGSX, SGGDX, USERX, VGPMX
Broad Commodities: DBC, DJP, GSG, GSP

It is important to acknowledge that Broad Basket Commodities were first seen in the New York Stock Exchange in 2006. As such, the models generated within this report do not precede mid 2006. CSV files were obtained for each fund including the date recorded for observation, the opening price, the high, the low, the closing price, the adjusted closing price, and the volume traded.

As specified in the methodology, a portfolio is generated randomly, and with given biweekly contribution and rebalancing period parameters, a weekly simulated history of the portfolio can be developed.

From this history, the standard Cumulative Annual Growth Rate (CAGR) is calculated to assess yearly returns, and the weekly record is evaluated to determine the worst total loss in 52 consecutive weeks.

5000 simulated portfolios were developed with randomized funds, biweekly contributions between 100 and 1000 dollars, and rebalancing periods that ranged from weekly to yearly (in 1 week increments).

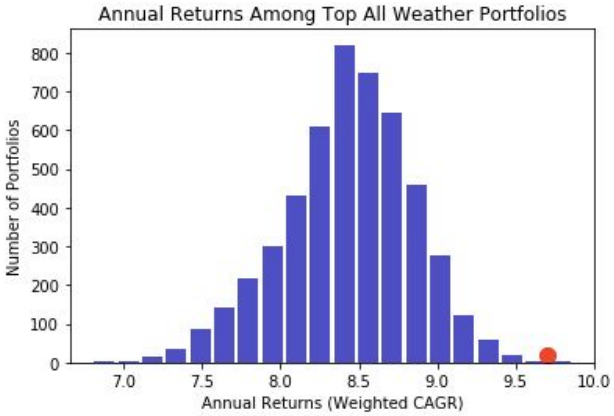
Accommodating the historical timeline

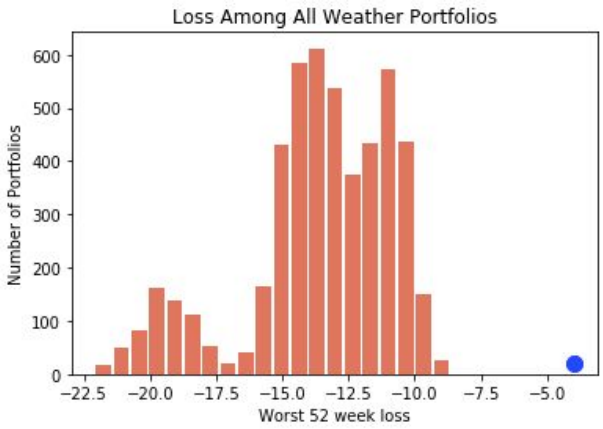
Since the stock market crashed in 2008, an alternate timeline was also crafted by reversing all the date from each daily observation. This placed the crash towards the end of the 11 year period, since this would provide another way of evaluating an accumulated portfolio's health under the stress of the 2008 recession. CAGR was calculated given the alternate timeline, and the deeper of the two losses was recorded as the worst year on record.

Statistics and Results

Our exploratory data analysis provided sufficient information to conduct an initial hypothesis test: An All weather portfolio should average a return of 9.7% Cumulative Annual Growth Rate and lose 4.3% in its worst year (reinvesting dividends).

Only 4% of random portfolios had an expected CAGR over 9.7 percent and 100% of the worst years calculated had losses exceeding 4.3 percent! The analysis implies that it is extremely unlikely the expected (mean) CAGR and loss values are 9.7 and 4.3 percent, respectively. Based on the analysis, it does not seem likely that this kind of portfolio performs as well as indicated.





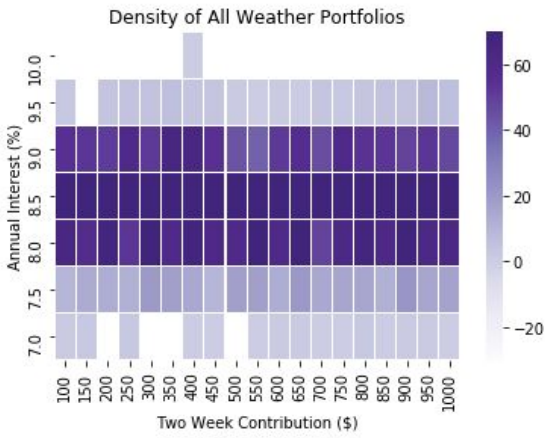
Considerations:

1. A different selection of funds within each category can create discrepancies in the final sample space. If a poorly performing fund was selected to contribute to the sample space, it will affect every portfolio in which it is sampled. However, the diversification and approachability for non-seasoned investors was the point of All-Weather.

2. The numbers hypothesized were formed by an analysis which dated back to the inception of the NYSE. While such comprehensive data is available for some of the funds in this data, it is not available for many. This study is limited in that its timeline only goes back as far as the inception of the first Broad Basket Commodities in 2006.

While the initial hypothesis does not seem to be well supported by the data, this exercise still provides some interesting results:

1. A bootstrap hypothesis test determined that there is no significant difference ($p = .20$) between those portfolios where the investor contributed $100 a pay period and those where the contribution was $1000.



Thus, any investor can take advantage of the kinds of returns an All Weather Portfolio can obtain.

2. Further examination of the results provide an idea of which funds and combinations of funds provide the best returns. Frequencies were examined for which funds appear in those portfolios with CAGR above 9 and which appear in those portfolios which experienced losses in the smallest quartile of all losses. As a result, the following optimally performing portfolio was recommended:

   STOCK: VIGRX, INTERMED BOND: IEF, LONG TERM BONG: WHOSX, GOLD: SGGDX, COMMODITY: No fund is significantly different from another fund. Rebalanced every 40 weeks

The recommended portfolio experienced a CAGR of 9.99%, with a worst year's loss of -9.33%. This is impressive, considering the normal losses experienced by the other All Weather portfolios - 10% to - 21%. How well does this recommended portfolio perform? The S&P's 10 year average through 2018 was 8.46%, and it's biggest loss in 2008 was 36.55%.

The portfolio most recommended by the data analysis delivers on the promise of keeping pace with the market and minimizing losses, albeit not as spectacularly as originally promised. It is a well balanced investment plan for an amateur investor. It is also worth noting that the software that has been developed to support this analysis can also be improved upon by considering alternate or additional categories of investment. There's no reason the funding for the stock index portion of the portfolio cannot be split up among stock indices and companies in promising sectors of the economy (like tech for example). Exactly how this should be done, and if such moves can be recommended by time series analysis is a matter for further study.

Tax consideration and fund fees have not been included in this model at this time
Past performance is not a guarantee of future performance