S&P Portfolio Optimization Project Report

Justin Bates
November 17, 2020

Problem Identification Overview

Context

The field of financial asset management is difficult to understand and navigate. It takes time and money to research and analyze assets. Some of the most exciting and high profile assets are assets in the form of securities. In the United States, financial securities are tradable assets such as debts, equities, and derivatives. Nowadays, the financial investment that most U.S. citizens think of is equities and, more specifically, common stocks. They have begun to believe it is the best indicator of economic growth.

Sure, you can buckle down and evaluate stock after stock, company after company, CEO after CEO, quarter after quarter. But is there an easier way for ordinary folks without the financial know-how or even time to make less risky investments? There is a way through the use of ETFs, Exchange-Traded Funds.

ETFs are portfolios that trade on exchanges just like any other stock. ETF portfolios can contain a combination of assets such as stocks, bonds, currencies, and commodities. ETF makes it much easier for an everyday investor to invest with lower risk and little to no supervision.

The most common ETFs follow the S&P 500 index (^GSPC) like The Vanguard Group (VOO), iShares (IVV), and State Street Corporation (SPY). The S&P is a capitalization-weighted index, which means it indexes the equities by allocating with capitalization, the stock's share price multiplied by the number of outstanding shares.

Problem

Can a portfolio optimizer focused on maximizing Sharpe ratio with current S&P 500 companies' adjusted close prices outperform the current S&P 500 index, ^GSPC, by 10% more cumulative return with a single year's data and the second year's data to view the performance of the optimized portfolio?

To generate a solution to the problem, we want to build an optimized portfolio from the current S&P 500 companies. To optimize a portfolio, we want to maximize returns while minimizing the risk concerning the amount of money we allocate on each asset in our portfolio. We want to think about index, sector, industry, and industry anti-correlation and covariance. We compare the portfolio to the efficient frontier. The efficient frontier is a set of optimal portfolios or assets that offer the highest expected return for the lowest risk.

How will we know if the optimized portfolio is more robust than the S&P 500 index, ^GSPC?

We will run a test on the subsequent year of data that the portfolio was built-on to see if the portfolio can hold more robust than the S&P 500 index, ^GSPC. We will be looking to reduce risk while also increasing return. The return should be 10% or more when adding little to no more risk.

Problem Solution

We outperformed the current S&P 500 index with a portfolio optimizer focused on maximizing the Sharpe ratio. There is some risk as the optimized portfolio's volatility is a little more considered because the return has higher peaks. We were able to test the portfolios using two different sampling methods

bootstrap and Gibbs sampling. We used bootstrap sampling to see distribution confidence intervals. While on the other hand, we used Gibbs sampling to generalize using a multivariate probability distribution, which caused us to obtain a much smaller sample of portfolio runs.

After evaluating the bootstrap sampler, we are 95% confident that the S&P 500 index's cumulative return is between 4.8% and 5.7%. The index has an okay mean cumulative return on investment of 5.27% for low-risk portfolio volatility of 0.23%. We are 95% confident that its cumulative return is between 26.7% and 28.7% for the optimized portfolio. The mean cumulative return for the optimized portfolio was 22.52%, with portfolio volatility of 0.47%. The return is much higher than the 10% increase. We observed with a 95% confidence between a 21.6% and 23.4% increase.

The Gibbs sampler gave us some different results as expected. We are 95% confident that the S&P 500 index's cumulative return ranges between -17.1% and 38.9%, and our optimized portfolio is 95% confident that its cumulative return is between -20.8% and 79.3%, with the Gibbs sampler. The Gibbs sampler shows different results than the bootstrap because the Gibbs sampler couldn't take as many samples. A Gibbs sampler using times series requires that the samples aren't correlated. The Gibbs sampler helps us evaluate if our optimizer might be overfitting to the dataset. We will review this further within the Optimization Model results section.

Data

I started by web scraping wikipedia.org/wiki/List_of_S&P_500_companies and slickcharts.com/sp500 for the current S&P 500 companies. I made comparisons to confirm their similarities and generate a final S&P 500 market-capitalization-weighted index. The datasets also gave me the corresponding GICS sector and sub-sector industries. I then used the python library, yahooquery, with built-in APIs to help query yahoo finance databases. I ended up only using the adjusted close and shares outstandings columns for each common stock ticker. We used this data to build optimization models and produce weights for the best new portfolio.

Deliverables

- Jupyter notebooks outlining the data science method used to produce models from dataset to solve the defined problem best
- A simple program that collects data builds the final optimizers and generates weights with optimizer models if you don't want to run through the Jupyter notebooks
- Project Report
- General Public Investors Summary Presentation

Data Preprocessing: Notable Steps

After using the yahooquery library, we had an abundance of data, way more than we needed for our optimization method. We only needed the stock symbols with their adjusted close with corresponding dates. We compiled this data into a pandas dataframe. Within this form, the data was nearly clean. However, two small transformations for analysis purposes still needed to be completed. First, We set the date columns to the dataframe index, and then the index of dates to datetime objects rather than strings. This step will help later with splicing. Second, We needed to identify any missing data. There were some missing data caused by us only using the current S&P 500 over 20 years even though the S&P 500

companies changed as an asset's market capitalization increases and decreases, allowing assets to move in and out of the top 500 companies. I left the S&P 500 as a constant but have plans to make it a changing index where assets can enter and please the index.

The results and intuition gained from the time series EDA step push me to take the 20 most extensive Sharpe ratio stocks and run them through an efficient frontier portfolio optimizer that outputs weights of the portfolio. This way, the optimizer doesn't need to use its resources evaluating the stock that it will eventually mark with a weight of 0. Figure 1 below shows the top 20 cumulative return assets' expected return vs. their risk throughout our twenty-one years of data, 1999 to 2019. Later in the Optimization Model Performances, Results, & Insights section, we will establish a specific year to compare the top 20 Sharpe ratio assets, the portfolio the optimizer builds, the efficient frontier, and the following return test year. Figure 2, on the next page, shows histograms of daily returns from 1999-2019. Twenty-one years of mean, standard deviation, skew, and kurtosis for the daily return. On the left of each histogram is a list of the top 5 Sharpe ratios and their corresponding assets. We could see what the S&P 500 assets' general trend was for the year with this information.

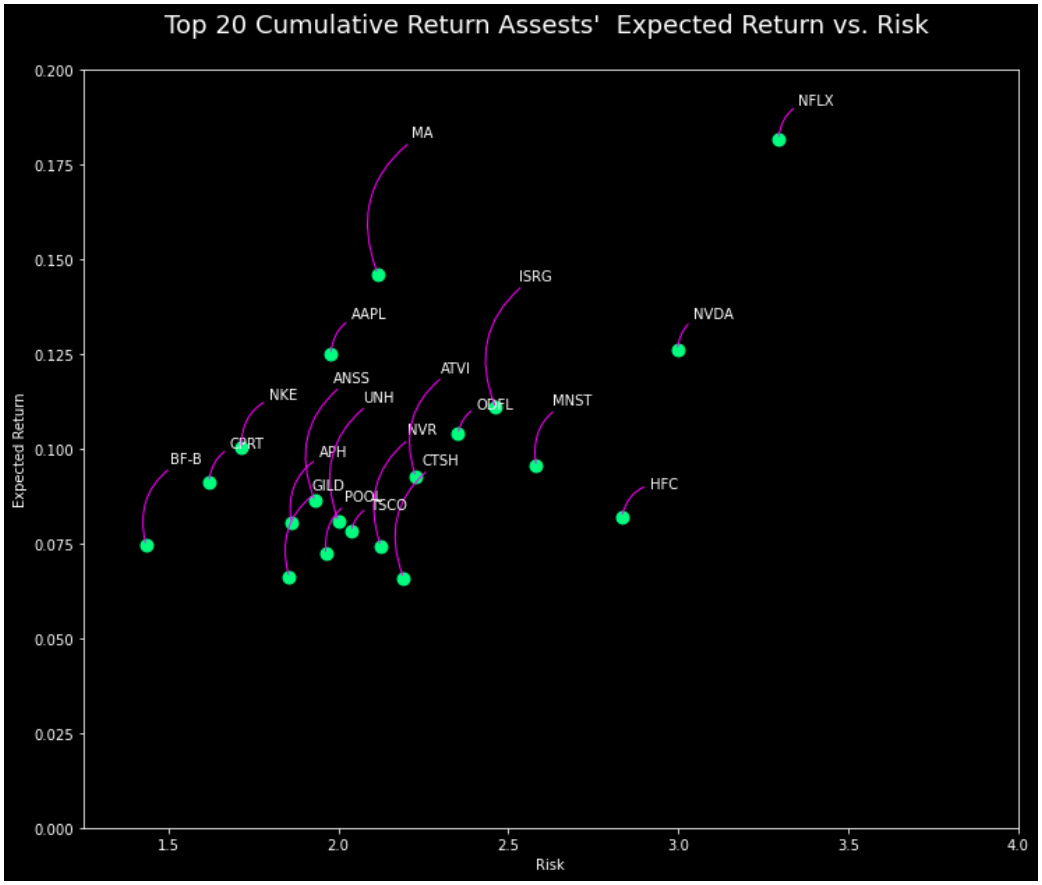

Figure 1: Top 20 Cumulative Return Assets' Expected Return vs. Risk

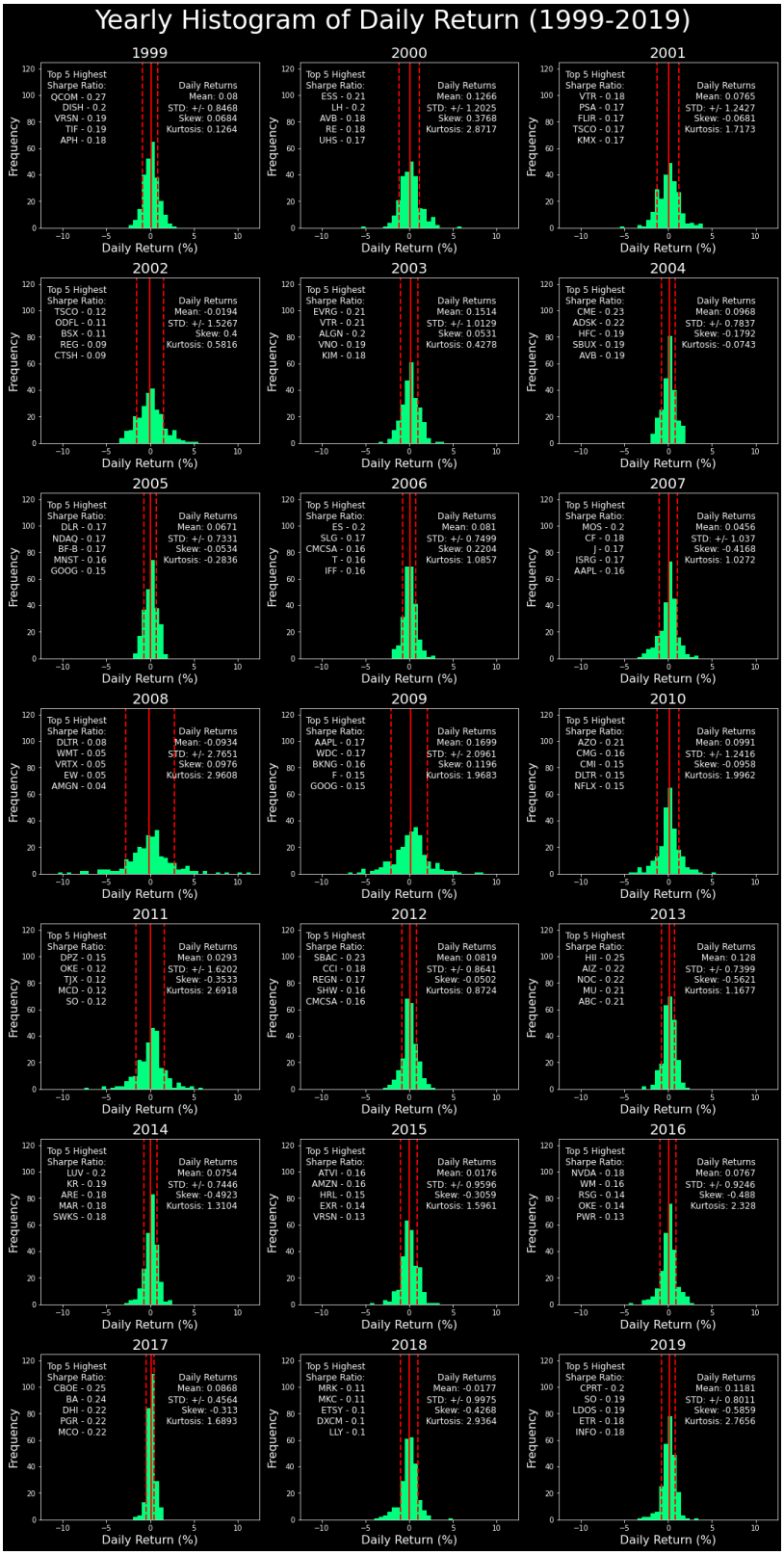

Figure 2: Yearly Histogram of Daily Return (1999-2019)

Optimization Model Description, Results, & Insights

Optimization Model Description

The efficient frontier portfolio optimizer requires an optimization objective and optimization constraints to solve the convex optimization problem. Our objective was to maximize return while using our maximize Sharpe ratio constraint. Our efficient frontier portfolio optimization model took in two inputs.

The first input was an expected return. Since we are using mean-variance optimization, the optimizer must know the daily return. These are practically impossible to know, but we can have some estimates by extrapolating historical expected return data. We found giving higher weights to more recent data with the exponentially-weighted mean of (daily) historical returns would produce a more accurate daily return with our portfolios.

The second input was the output of a risk model. The risk models we played with were sample covariance and semi-covariance. We found a semi-covariance risk model, a measure of all returns below some benchmark variance, works best to diversify the portfolios' different assets. We used our risk-free rate of 0 to set the benchmark. This strategy creates a measure of downside risk in our return.

Figure 3 below shows the semi-covariance risk model and correlation with a heatmap for the year 2016 with the top 20 Sharpe ratio assets. We will be using this year of data to convey insights of the optimizer in the Optimization Model Insights section. All of these top 20 Sharpe ratio assets have nearly 0 covariances. Since we are using the semi-covariance, we can see that this only allows for a positive correlation. We decided on this strategy because our portfolio won't contain any shorts. Our optimizer is defined such that it will only take positive asset positions, no shorts. A short, or short-selling, is a negative position strategy of borrowing shares, immediately selling them, and hoping to repurchase them at a lower price to return to the lender while keeping the difference as profit.

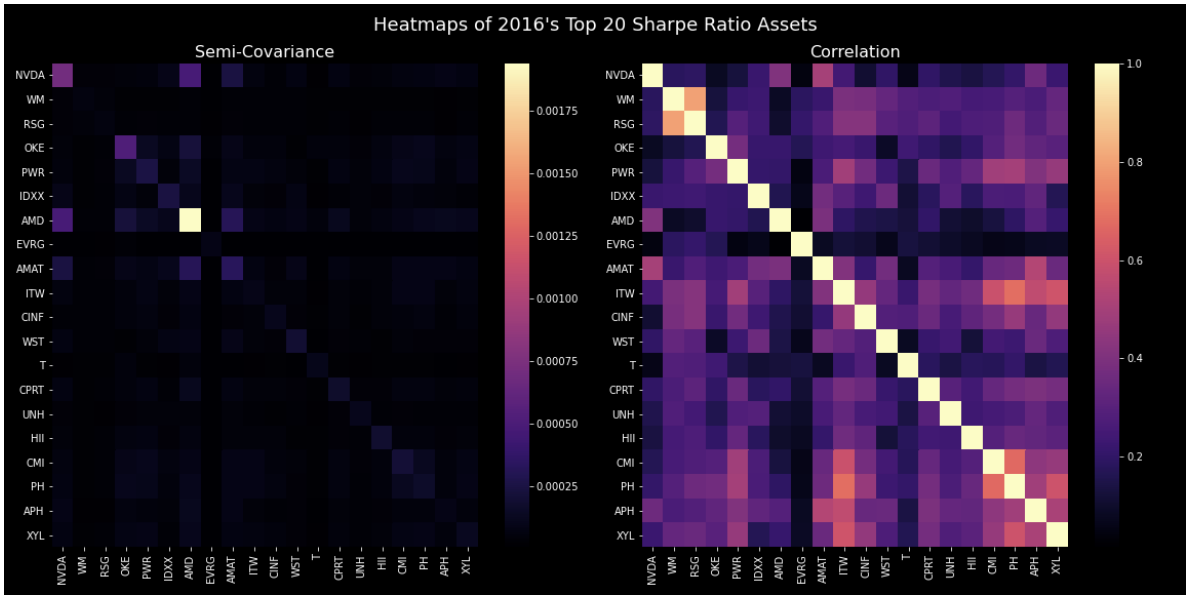

Figure 3: Top 20 Sharpe Assets Semi-Covariance Heat Map for 2017

Optimization Model Results

As stated previously in the Problem Solution section, we started to evaluate our optimization model using bootstrap sampling to see the distribution differences between the S&P 500 index and our portfolio generated using our optimization model. The bootstrap sampler took 300 samples with 75% of the dataset. We used Gibbs sampling to develop an intuition on how well the optimization model performance holds with the distribution only containing samples that aren't correlated. The Gibbs sampler was able to retrieve 18 uncorrelated samples.

Figure 4 shows the cumulative return distribution from bootstrap sampling for both the max Sharpe optimized portfolio and the S&P 500 index. We can right away see the difference in spread between the two portfolios. Using the bootstrap sampler, we are 95% confident that the optimized portfolio is between 26.7% and 28.7%, and the S&P 500 index is between 4.8% and 5.7%. The mean of the optimized portfolio is 27.79%, while the volatility is 0.54%. The S&P 500 index was much lower, at 5.27% for the mean cumulative return and 0.23 for the volatility. Figure 5 illustrates the difference between the two asset portfolio's cumulative return. The goal set out was 10% greater, but the difference in the two is between 21.6% to 23.4% with a 95% confidence. That is more than double our goal.

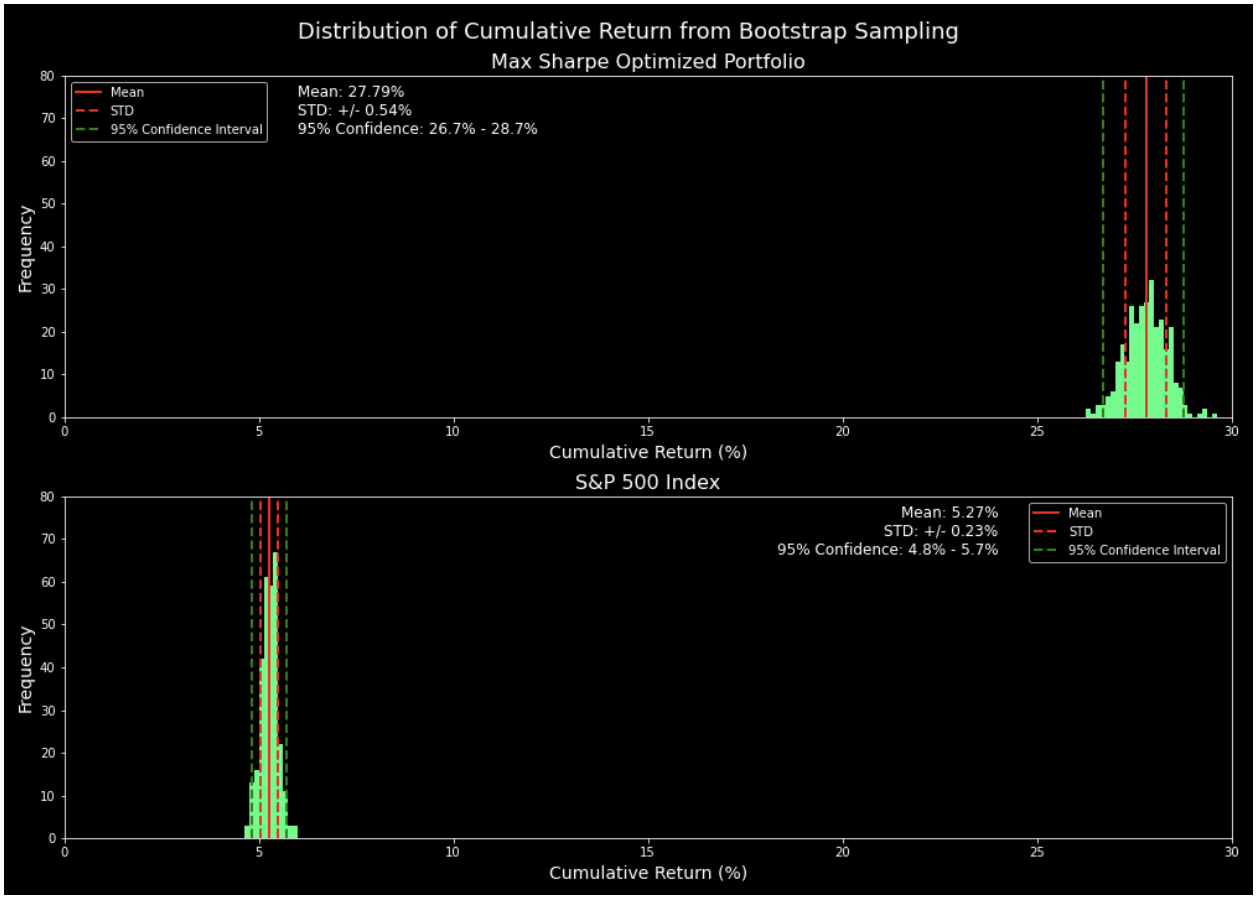

Figure 4: Distribution of Cumulative Return from Bootstrap Sampling

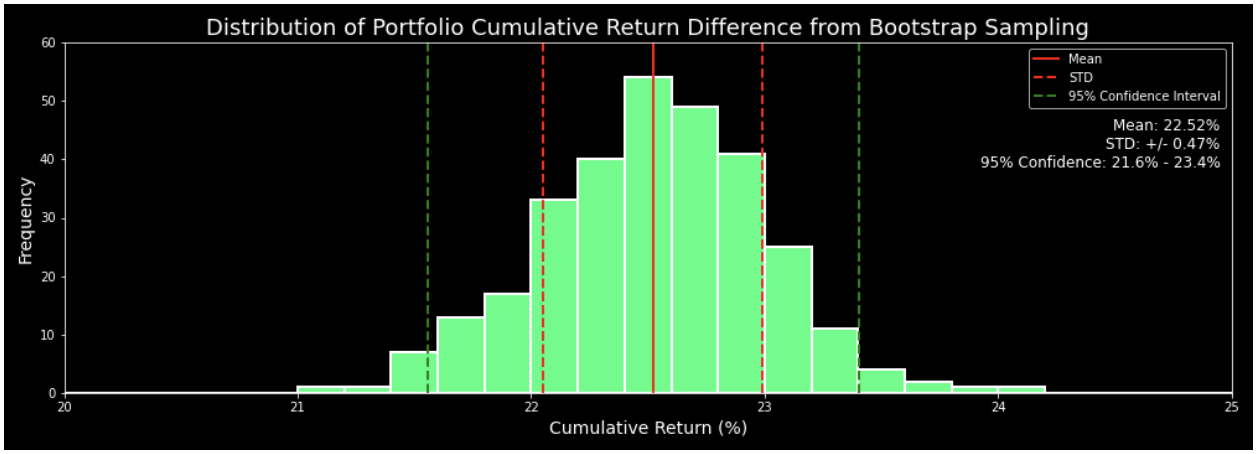

Figure 5: Distribution of Portfolio Cumulative Return Difference from Bootstrap Sampling

Figure 6 reveals the second sampling method used, Gibbs sampling, with both portfolios' distributions of their cumulative return as well as their differences, similar to the bootstrap sampler previously discussed. With these two distributions, we can see some similarities as the volatilities are much larger. Using the Gibbs sampler, we are 95% confident that the optimized portfolio is between -20.8% and 79.3%, while we are 95% confident the S&P 500 index is between -17.1% and 38.9%. The mean of the optimized portfolio is 28.0%, while the volatility is 28.0%. The S&P 500 index was much lower at 9.19% for the mean cumulative return and 15.5% for the volatility. Figure 7 draws upon the difference between the two asset portfolio's cumulative return. Again, we set the goal of 10%, but the optimized portfolio and S&P 500 are less reliable in this sampling method. We are 95% confident that there is a cumulative return between -21.6% and 81.1%. The return spread is way too volatile at 29.97%.

We get these results with large volatility because the model over fits the dataset. We can forgive some overfitting since the market is tough to generate expected returns using historical data.

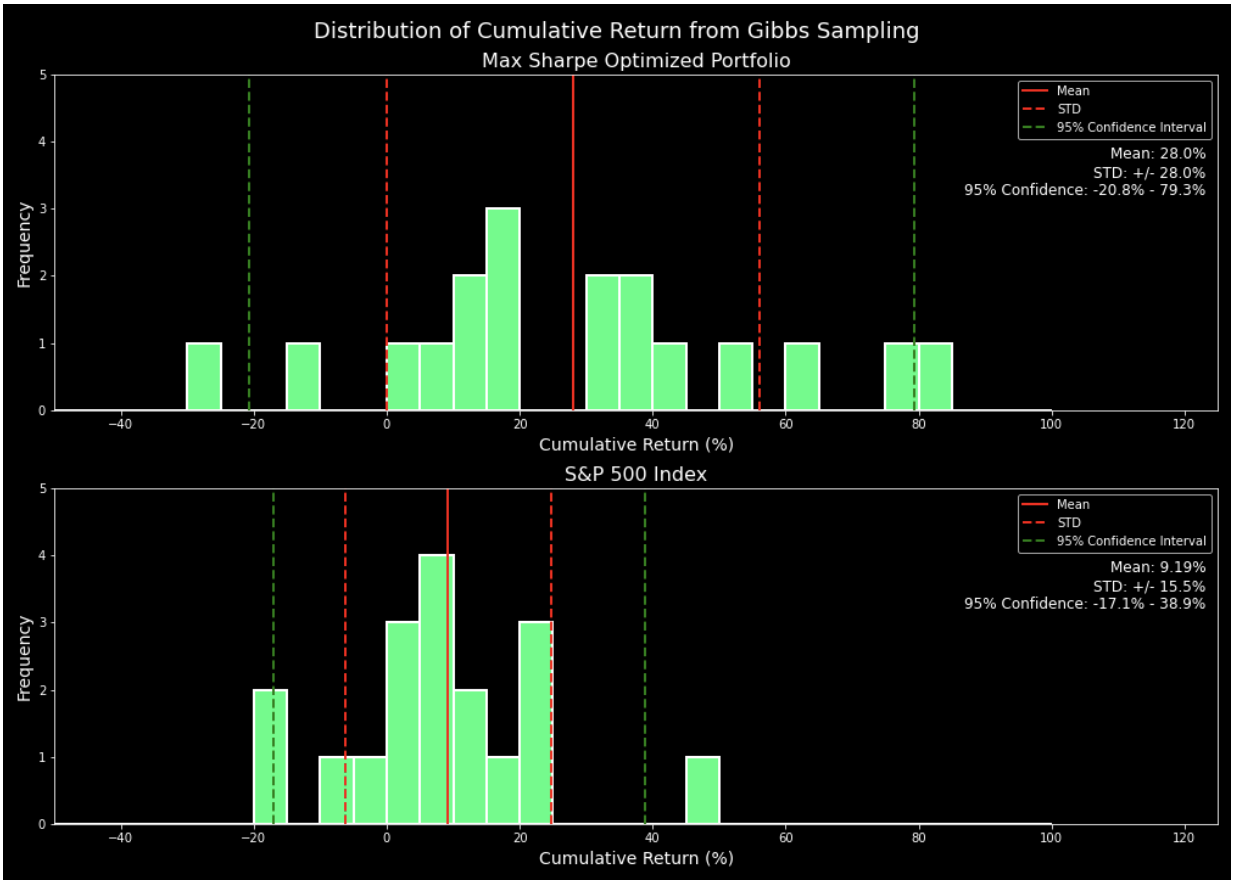

Figure 6: Histogram of Cumulative Return from Gibbs Sampling

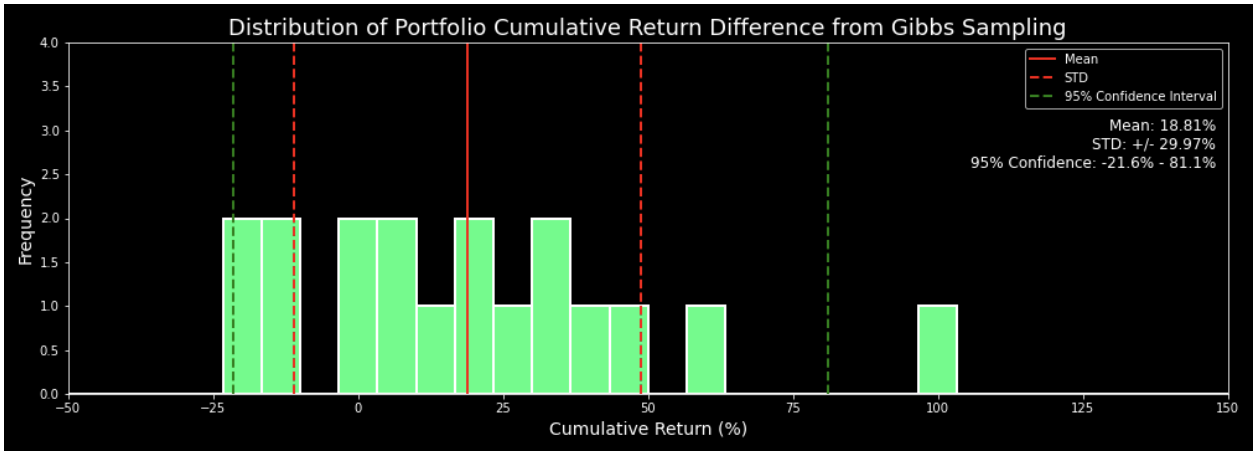

Figure 7: Histogram of Portfolio Cumulative Return Difference from Gibbs Sampling

Optimization Model Insights

We want to use our portfolio optimization model to find insights. Now let's look at a single sample as an example that takes a range starting January 1st, 2016 and ending December 31st, 2016. Our model takes in two inputs, the expected returns calculated by the exponential moving average historical return and the risk model, which estimates the semi-covariance matrix, shown previously for this example in Figure 3. Before passing the S&P 500 assets, we evaluate the Sharpe ratio for that time and remove all assets but except the top 20 Sharpe ratio. We take the historical adjusted return data for those top 20 assets and pass it to our model.

Figure 8 shows our optimized portfolio over one year of evaluation, 2017, has a cumulative return of 48.09% compared to the S%P 500 index's 18.41% return. We can see apparent volatility in the optimized portfolio compared to the S%P 500 index.

Figure 9 shows the efficient frontier for the current example of 2016 20 top Sharpe ratio data. We can see all return vs. risk for all 20 assets. The assets with weights greater than 0 are marked this a gold point while the remaining assets have silver points. The blue point identifies the max Sharpe ratio when choosing random weights for all 20 assets. The blue point's volatility is 0.04, while the return is 0.13, creating a Sharpe ratio of 0.31. The green point conveys a max Sharpe 0.22 for when choosing random weights for the gold points, weights great than 0. Lastly, the red star reveals the optimized portfolio on the efficient frontier. The Sharpe ratio of our optimized portfolio is 0.2. Our portfolio should be as close to the edge of the efficient frontier in theory, but it is impossible in practice. The efficient frontier's assumptions don't correctly represent the real world.

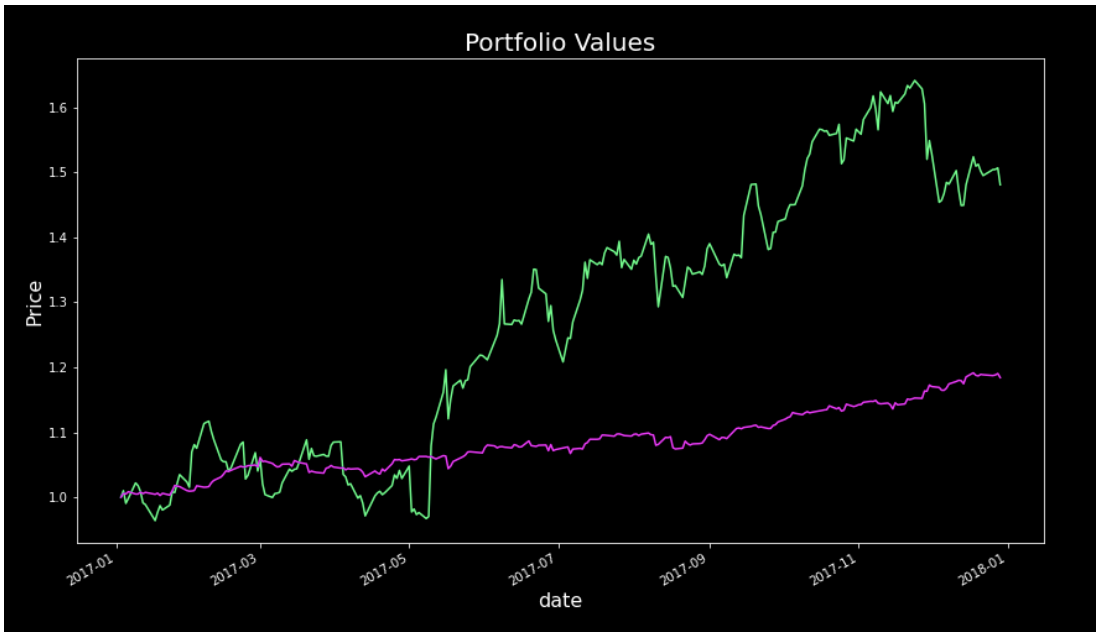

Figure 8: Optimized Portfolio and S&P 500 Index 2017 Cumulative Return from 2016 data

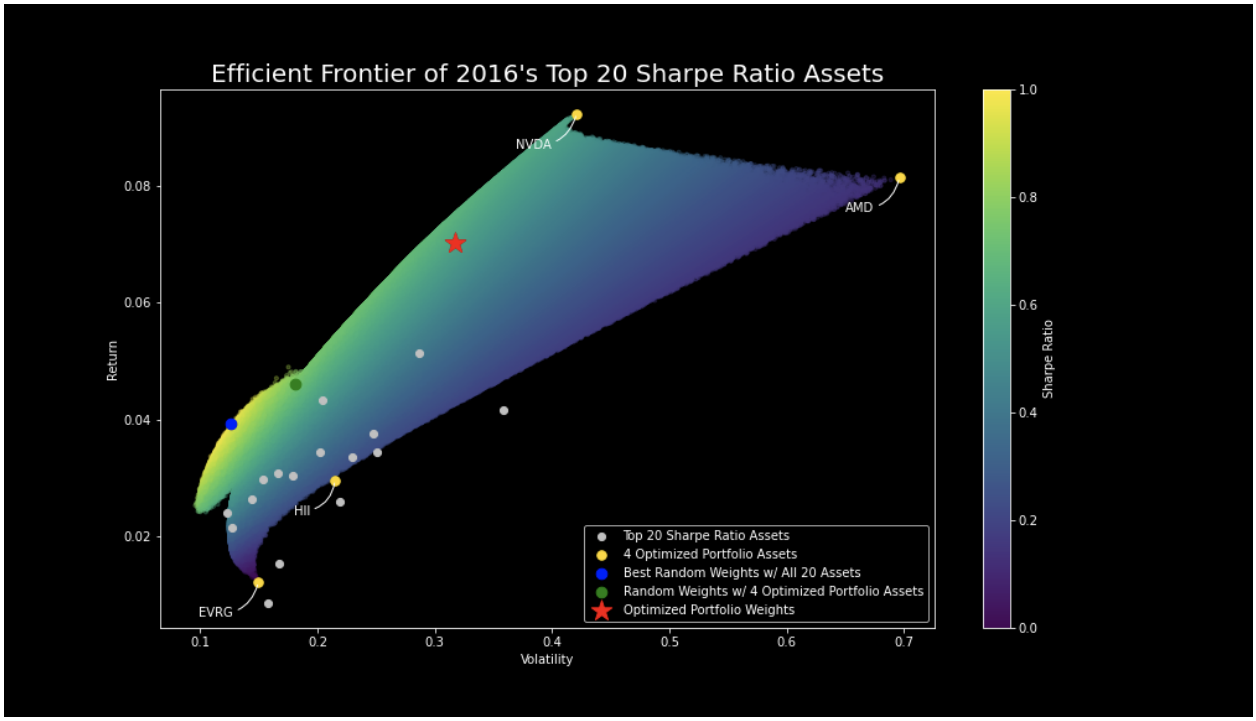

Figure 9: Efficient Frontier Assets and Portfolio for 2016 data

Conclusion

To conclude the report, portfolio optimization is a great tool that can help optimize portfolio objectives given a constraint. We can use the maximum Sharpe ratio as a constraint while optimizing for an objective of maximum return. This strategy may be a little overwhelming for an everyday investor, but it could be a simple way to build an ETF that can produce returns more significant than the S&P 500.

Next Steps

1. Use the S&P index stocks to evaluate that given period rather than using a fixed set of the current 500 stocks.
2. If I were to use the S&P stocks for their given period, I could use a larger dataset to evaluate my results.
3. I would like to build an optimizer from scratch using deep learning methods.
4. Use a similar strategy to rebalance potential portfolios.
5. See the effects of a different train, test, and evaluation durations.
    a. Could we see a larger return if we use a swing trading strategy with a month of data and re-evaluate each month?
    b. Could we see a larger return if we use a long hold trading strategy with three years of training data and re-evaluate each year vs. using just one year of training data as we did in this experiment?