**Get Your Hands off My Money! – A Machine Learning Approach to Asset Allocation**

In the 1950’s Harry Markowitz introduced the financial world to Modern Portfolio Theory (MPT). The engine behind modern portfolio theory is called mean variance optimization (MVO). MVO is a fairly simple optimization problem whereby a practitioner seeks to identify efficient asset allocations based on expected returns (mean), and risk (variance), subject to any number of constraints about position size, long/short rules, leverage and others. The theoretical importance of this idea was strong enough to win Markowitz a Nobel Prize in 1990. However, MPT in its original form has one glaring problem…its not particularly applicable to the real world. To be clear, the majority of technical issues that MPT presents for real world practitioners are manageable, it’s the theoretical ones that are more problematic. Particularly problematic is the use of expected values for means, and variances/covariances, after all there is no crystal ball.

The following document is divided into two sections. The first section will focus on supervised machine learning methods, well the second will focus on unsupervised methods. Both sections will include a brief discussion of the features used in the experiment a description of the machine learning methods used, and an evaluation of outcomes.

Process

The following document is divided into two sections. The first section will focus on supervised machine learning methods, well the second will focus on unsupervised methods. Both sections will include a brief discussion of the features used in the experiment a description of the machine learning methods used, and an evaluation of outcomes.

Supervise methods.

Is is often said that trying to time the stock market is a fools errand. In fact there is a very well-known theory in the financial world that describes the stock market as a random walk. The idea behind the random walk hypothesis is simple. Asset prices evolve randomly and therefore cannot be predicted. Assuming for a moment this theory holds, it begs the question, “what exactly or financial professionals doing all day?” at some point in the seasoned asset manager comes to the conclusion that investing is not about prediction, it’s about probabilities. Or is the data science community would put it it is a classification problem not a prediction problem. The first model i present deals with precisely that matter. Formally, The model seeks to answer one question. Giving the various macro economic data and market data available at any given time is it possible to accurately assess the probabilities of adverse events in the financial markets?

Data and features

1. Shiller’s‘s PE ratio. Also known as the cyclically adjusted Price to earnings ratio. This multiple is used to gauge value in the stock market, it is similar to the very common price to earnings ratio. The difference is that the price to earnings ratio generally uses trailing 12 months of earnings, where is Shiller’s’s PE uses 10 years of earnings adjusted for inflation as a gauge. The idea being if you economy works in cycles, in short term earnings can be very noisy, cyclic we adjusted earnings over the span of 10 years or more stable and provide a stronger indication regarding a firm/markets earnings power.
2. Yield curve slope. The slope of the yield curve is calculated by subtracting videos of a three month United States treasury bill from the yield of a United States 10 year treasury bond. There is historically been strong empirical evidence supporting the idea The slope of the yield curve alone is a very powerful predictor of looming recessions. Well there is no clear consensus on why or when it yield curve inverts, or even how long short term meals must be above longer-term yield for it to be considered inverted, some theories exist. Among them is the idea that market practitioners that expecting The federal reserve to lower rates in order to combat a slowing economy, attempt to Lock in attractive yields by buying longer duration bonds ahead of interest-rate cuts. This drives the price of the bonds up and the yield of the bonds down below them work of the three month bills.
3. Core inflation as measured by the Consumer Price Index (CPI). The Consumer Price index is a measure of inflation. This index is released each month by the US Bureau of Labor Statistics. It is meant to reflect the changes in the prices of goods and services purchased by consumers in the United States. That being said The standard Consumer Price Index has some problems. In particular it includes a few items that have very volatile prices. These items include food and energy. To deal with this the BLS also put out the Core Consumer Price Index that excludes both these categories and provides more more stable and accurate indication of inflation.
4. Unemployment: Unemployment is for the most part considered a lagging indicator. The highly publicized non-form payrolls come out on the first Friday of every month and provide The public information about the month before. In other words, market practitioners and economist generally don’t expect to see unemployment rise until after trouble has started. That being said, The theory of full employment clearly states that structural on frictional unemployment may remain even when the economy is at for employment, in fact the OECD believes that the full employment unemployment rate in the United States ranges between about four and 6 1/2% . Given the fact that economies are cyclical, One could logic we look at the inverse of the unemployment rate as a leading indicator. In other words, when employment calculated as one minus the unemployment rate reaches levels of 95%+ it could be an indication that economy has in fact reached a peak.
5. SM purchasing managers index. This is often the first major piece of information economists and market practitioners receive every month regarding the health of the economy. The index is based on the survey conducted monthly by the Institute for supply management. The survey includes questions regarding production hiring supplies deliveries and inventories. The index is adjusted for seasonality. A PMI above 50 indicates expansion in the manufacturing segment of the economy below 50 indicates a contraction.
6. Personal income and outlays. This Numbers reported monthly by the Bureau of economic analysis and tracks consumer income and spending. Historically household consumption which represents the C in the well known C (consumption) + G (Government spending) + I (Investment) +[X-M] (Exports - imports/ net exports) formula for gross domestic product, average is about 68% in the United States. As such we expect that this metric can provide insight into a slowing economy.
7. The University of Michigan consumer sentiment index. This index is a measure of consumer confidence. It is published month by the University of Michigan it’s stated objectives are: to assess near term consumer attitudes on the business climate personal finance in spending, to gauge the economic expectations in probable future spending habits of consumer, to judge the consumers level of optimism/pessimism, to promote an understanding of into forecast changes in the economy. The monthly survey consists of 50 core questions and includes at least 500 telephone interviews across the contiguous United States (excludes Alaska Hawaii and other US territories).
8. The Vanguard 500 stock index fund The Vanguard 500 stuck index fund is the first of its kind and the oldest passive investment vehicle still active. The fun begin trading in August 1976 and currently has About 637 billion dollars in assets under management. This is incorporate in the model as a vehicle for investment in the broad stock market.
9. The Vanguard total bond market index fund- also the first of its kind this fund provides exposure to the broad US bond market it begin trading in December 19 86 and is incorporated in the model as V means by which an investor would invest in the bond market.
10. NBER business cycle (Model Label)- The data used to create the label for the model is published monthly by the national Bureau of economic research. It is a binary index. If the economy is an expansion the index Re:ZERO, if the economy is in recession the index reads one.

Future preparation

Access

All the Raw data was imported via the data stream web services API. The request asks for since inception data using monthly intervals for all macro and financial market information. Fish data is stored in CSV file for further use.

A second call to the data stream web services API imports The total return index for both Vanguard funds. this file two is stored in a CSV file for for the use.

Processing

Label: Tomorrow must be forward-looking to be of any value. An indicator that accurately identifies that the economy is in recession two months after the recession has started is of no value. To ensure that our model is forward-looking we shift our labels back by 12 months. For example We now know that do United States economy officially entered a recession in February 2020, this would be represented in our Rod data as with a one

By shifting all observations back 12 months and marking February 2019 with a one instead, Our model will now evaluate conditions a year before a known recession.

Features transformations

Rolling normalizations

Constructing an indicator using date it it has been normalized over the entirety of a data sets life is a little value. Consider for a moment one of our features Shiller’s‘s PE ratio. This indicator reached an all-time high in December 1999 during [the.com](http://the.com/) bubble. The issue is that every month between April 1998 in December of 1999 was also a new high. Standard normalization across the entire day to set would not capture this properly Because when the information was in fact reported A practitioner would’ve drawn conclusions based on each of those months being the high. As such every month as new information is reported the mean of the information changes the volatility changes and every observations standard score changes. To account for this, The model transforms the data using And expanding normalization process with a 36 month minimum observations window. This process works as follows:

1. calculate expanding mean
2. Calculate expanding Standard deviation
3. Subtract a given observation from its corresponding expanding mean and normalize by the expanding standard deviation

Train test split

We run the models features and label through the standard SK learn train test split function using a test size perimeter of one one half.

Our classification model of choice for this task is the gradient boosted decision tree classifier. This model copiously been built using a support factor machine classifier or a logistic regression, The decision to use The gradient boosted classifier came down to the fact that it does not require careful normalization of features to perform well, and given the less than conventional normalization process employed in the feature preparation The gradient boosted classifier was the most prudent choice.

Model training

Gradient boosted classifiers are sensitive to two parameters, The first is the number of estimates The model will use in the second is the learning rate. The number of estimators corresponds to the number of small decision trees used in the model, whereas the learning rate controls the emphasis on fixing errors from One decision tree to the next.

The model selection process:

1. Establish the number of estimates to test in advance, and this model we evaluated models with 100, 200, 300, 400, and 500 estimates
2. Establish a range of potential learning rates. And this model we evaluated 0.0001, 0.001, 0.01, and 0.1
3. Train a model for each combination of learning rates and estimates
4. Calculate micro F 1 scores for each model Using test data And record
5. Select The model with the highest F1 micro score.
6. Once model parameters are selected we use the cross validation score function in SK learn using three cross validations in the same F1 micro scoring parameter. To ensure consistency. We choose three cross validations simply because recessions are not very common and do not last for very long, as such a high number of cross validations may result in sampling problems.

Once the model is trained, fit, and applied to the test data, we use the predict probabilities method to extract the probabilities of any of our test labels in classified as a 1 (expect a recession in the coming 12 months) or zero.

Discussion of results

Model scores

Used used F1 as an evaluation statistic simply because Minimizing false negatives and false positives or more important than the overall accuracy of the model. It’s important to remember that The supervise learning model in this case is being used to make decisions about portfolio allocation, as any season portfolio manager knows avoiding mistakes is exponentially more important then making good calls (and no they are not the same thing).

Using the standard train to split The F1 micro score of our selected model came in at 0.93984

Our cross validation evaluation using three cross validations and F1 micro scoring produced in average F1 micro score of 0.8496 and a standard deviation of 0.0278.

Feature importance

Not surprisingly, The slope of the yield curve posted the strongest feature important to 29%. This was followed by valuation 17%, inflation personal consumption and purchase managers index we’re fairly similar with each between 13% and 11%, and employment in the University of Michigan sentiment indicator both came in at 8%.

The results observed make sense. If we evaluate the strongest two features in terms of feature importance (slope of the curve and valuation) we know if they both have to do with financial markets. And both are forward-looking mechanisms that discount future expectations to today’s terms. In other words, both of these features are constantly telling us in real time what to expect in the future. As such, when these two features reach certain levels implications for the future given what is currently known at any given point in time are baked in.

Real world application.

When we first plot the indicator and overlay recessions since 1987 on top we notice a few issues. First moves in the raw indicator are very sharp probability of a recession in the next 12 months goes from extremely low to extremely high very very quickly. To to adjust for this, we take the exponentially weighted moving average of the raw indicator with a span decay of four. This helps smoothing out some of the sharp moves the raw indicator has.

Second we notice that Wyler indicator does a very good job telling us that a recession is coming When a recession is 12 months away, as time goes on the model begins to pick up a recovery often times before the expected recession even begins. This is likely the result of the fact if the average recession lasts anywhere between six and nine months, whereas the average expansion lasts years. Unfortunately this could pose a real problem for portfolio managers attempting to use this indicator as it stands For the simple fact that sell signals will occur early and in advance of a recession as desired, but subsequent buy signals will occur as the recession is starting A head of the desired entry point. To be clear this is not a problem with the model rather It dictates how the model is to be used. As such, we chose to deal with this issue in the asset allocation function itself.

Asset allocation

That’s far we’ve taken financial market in macro economic data cleaned it prepared it and run it through a classification model. The exponentially waited moving average of The output from that model acts as an indicator that a portfolio manager could use to make changes to their asset allocation. While the data has established that the indicator can provide insight into The probability of a recession, we have yet to establish the quality of decisions made based on the indicator itself. To that end, The next step in our process is to simulate portfolio performance using a rules-based process for allocating capital that revolves around our indicator.

Acid allocation process

To run our simulation, we will create a simple portfolio that consists of stocks and bonds. We use index funds to eliminate any undesired affect from security selection or sector allocation.

The allocation process works as follows:

1. start with a standard 60% stock 40% bond portfolio
2. If our indicator is at or above 0.7 at any given time over the past six months, We change the allocation from 6040 stock and bond to 4060 stock and bond.
3. Once the indicator has not read a level of 0.7 or higher for six months we move back to 6040 stock.
4. The allocation is shares based, as such we make no changes to the portfolio as long as there is no signal from the indicator. This means that shares are calculated the moment The indicator gives us signal signal, and the number of shares held his constant until the next signal

We ran our back test using monthly intervals between November 19 87 in December 2020. To evaluate performance, we looked at two metrics first and probably most obvious is performance over time in comparison to a 6040 buy-and-hold portfolio. A 6040 buy-and-hold portfolio is simply a portfolio that allocates 60% to Stocks in 40% to Bonds and just holds the same amount of shares in both without changes or rebalancing. We found that over the period in question the portfolio created using the indicator had a total return of 2272.12%, where is the buy-and-hold portfolio only posted a 2051.7% total return. In addition to absolute performance we evaluated the portfolio performance using Sharpe ratio. The Sharpe ratio is probably the most common metric used to evaluate per folio performance on a risk-adjusted basis. Calculation of the shower for ratio is as follows:

The average annual return of a portfolio minus the average risk-free rate divided by the portfolios standard deviation.

The intuition behind the sharp ratio is straightforward. Investors are interested in maximizing return for every unit of risk-takin. As such by isolating only the portion of returns coming from risk and then normalizing it by the portfolios volatility which is a proxy for risk, investor can understand how much return for every unit of risk here she is getting. As such the higher the sharper ratio the better. In our case The portfolio constructed using our gradient boosted decision tree indicator achieves a portfolio sharp of 0.91 where is the 6040 buy-and-hold portfolio only achieved a sharp ratio of 0.69 indicating that the model constructed here appears to add value over a 6040 by and hold approach.

Unsupervised Methods

Risk parity

Perhaps one of the great greatest challenges in the quantitative portfolio management is the delicate balance between the need to be forward-looking in the complexity that forward-looking models introduce. Making good predictions about future asset prices is very difficult, practitioners are never operating with all information, market participants do not act in a rational way, and exogenous events expected to occur once every hundred years seem to happen every 7 to 8 instead. Even if one could make strong predictions about asset prices, this is not enough. Equally important and exponentially more difficult is making predictions about covariances of assets. Using Standard econometric message attempting to project joint distributions of multiple assets into the future is an extremely difficult task, and often times requirements are relaxed and stationary and normality are simply assumed. This often leads to less than optimal results. As it happens principal components analysis can help deal with these issues.

Principal components analysis

This model will seem to take it vantage of tutee features of principle components analysis.

1. dimensionality reduction- The ability to deal with less data without sacrificing much from the results is very beneficial as it allows us to focus on information that is truly impactful on our decision making process.
2. Principle components are by definition orthogonal.- The very process of principal components analysis results in data sets that are by definition not correlated to each other. This illuminates the need to project covariances when doing forward-looking work.

This model, Will take a series of asset prices from various asset classes, decompose their returns into principal components, select all principle components that cumulatively explain 95% of variance among asset classes, project the volatility of each of those principal components over a given time horizon, reconstruct a covariance matrix using projections made from principal components volatility, in construct a risk parity portfolio using the forward looking covariance matrix generated using the principal components.

Features

We use daily prices from the following exchange traded funds as proxies for various asset classes in the model.

< insert list here>

Access

We access All data for the ETFs above, The other affinitive Eikon API. The data is imported and stored in a CSV file for later use.

Adjustments to features.

All prices are converted to percentages and any not a number observations that results from the process are replaced with zero.

Principal component selection.

We employed the following method to select the number of principal components included:

1. Train and Fit a principal components analysis on all the data.
2. Evaluate the explained variance ratio for the results
3. Select the minimum number of principal components required to explain 95% of variance among features.
4. Train and fit a new principal components analysis based on The established number of principal components.

Volatility projection

Once the principal components analysis it’s complete, we can proceed with volatility projections. To do so we employ the generalized autoregressive conditional Heteroscedasticity process or GARCH process. The gorge process was developed by an economist named Robert Engle in 1982, angle won the 2003 Nobel prize for economics for this contribution. The underlying principle behind the garbage process is that unlike a linear process when dealing with volatility air terms are not assumed constant. This makes any attempt to project volatility using linear methods problematic. The formula for a gorge process is as follows:

< insert>

This formula is fairly straightforward to implement in python via the arch package.

Once we have projected The volatility of our principal components, we must now reconstruct the covariance matrix using this information.

To do so, we take the following steps:

1. take forward to any projections for each principal component and arrange them in a diagonal matrix (Delta)
2. Multiply The transpose principal components loadings Matrix (Alpha)
3. Multiply the The product of alpha transpose and delta by the matrix alpha.

This process (orthogonal garch) was developed by Professor Carol Alexander at the University of Sussex and is detailed a great length in her book market models a guide to financial data analysis.

We now have a forward-looking covariance matrix and are ready to generate an asset allocation. To do sell we will use the risk parity asset allocation process.

Risk parity

For the most part modern asset allocation techniques are a simple optimization problem. We have a vector of returns, and a covariance matrix and we are in fact trying to either minimize variance, maximize return, or maximize return per unit of risk. And we do this by changing the weights assigned to each asset.

Risk parity is a variant of the above described process except that The focus on risk parity is risk budgeting. The idea is to ensure that assets are contributing to risk equally.

Do you vantage of using the risk parity model for this particular situation is that risk parity is entirely independent of expected return inputs. All that matters is the covariance matrix and risk budget we establish. This makes the model the perfect candidate to evaluate the covariance model we created using our unsupervised machine learning method.

The risk parity portfolio construction process works as follows:

1. we begin with an arbitrarily weighted portfolio.
2. Calculate the portfolios volatility this is calculated as...
3. Calculate each assets contribution to risk
4. Calculate the error term based on these results
5. Adjust weights to attempt to improve on previous iteration

In python we used the SCiPY stats optimizer minimize function to Execute this process. We also included in equality constraint to ensure that the sum of weights equal one, and we include in any quality constraint to ensure no short selling.

Discussion of results

Principal components analysis principle components analysis

The data set includes 12 features all of which are financial instruments that tree daily. As such we expected a relatively high degree of explanatory power from principal components. In fact the first principal component explains 68.3% of the variance among assets The next component explains about 13% of the data it takes the next six principal components combine to reach the 97% threshold. While it is impossible to tell exactly what each of the principal components are, it is highly likely but the first one represents the stock market. The second appear to be various components of the bond market including interest rates.

Portfolio construction

To evaluate the process overtime we once again ran a back test using the following process:

1. Portfolio trades on pre-establish dates in this case the 15th of every March, June, September, and December. Should stay stay following weekend The closest trading day is selected.
2. On every trading we train and fit a principal components analysis with daily returns for all assets over the previous year.
3. We extract the transformed data and project 30 days of volatility for every one of the principal components via the gorge 11 process.
4. We reconstruct the covariance matrix with the results from the gorge process and the components extracted from the PCA.
5. At this point we apply the risk parity optimization function to The new covariance matrix.
6. Once the risk parity function provides us with weights we rebalance our portfolio accordingly.

Evaluation of results

Machine learning results.

The principal components analysis during each iteration performed as expected. Whenever employing principle components analysis our expectation is that we end up with

a) a data set that is smaller and easier to deal with without sacrificing too much information

B) that all the principal components are not correlated to each other

Given the fact that we did manage to reduce the Data sets and the resulting data set allowed us to quickly and easily project volatility into the future it is clear that this method is very effective for such purposes given the Notion that the alternative is far more time-consuming, complex, and does not ensure better results.

Evaluation of real world application

Just as we did in section 1 we evaluated our process using The same methodology, The total return of the portfolio over the period in question , The portfolio Sharpe ratio. Unfortunately in this case, our simulation produced less than desirable results. I believe this is attributable to the following factors.

1. The inclusion of low volatility asset classes like short term bonds is a cash proxy, and United States treasury‘s as well as investment grade corporate bonds Will create a constant drag on portfolio performance.
2. During periods of of high volatility like Q1 of 2020 it is impossible for assets like equities and commodities to contribute the same level of risk to the portfolio as lower volatility assets without the use of leverage

Evaluation of acid allocation over time using this approach demonstrates both issues clearly, short term US treasuries are always a constant and usually above 20% of the portfolio, as our US treasuries, investment grade corporate bonds are constant until Q1 of 2020 where the model kicks them else increases exposure to US treasuries this exposure stays consistent for the remainder of the year. On a risk-adjusted basis the results are similar as the portfolios Sharpe ratio comes in at about 0.41 while a bond portfolio comes in at 0.48 in MCI or country index comes in at 0.48 as well.

Potential improvements to the model include changing the mix of acid classes, more frequent trading periods, paeans and potentially different optimization processes. It is important to note that all the changes required stem from domain issues, not from machine learning/data science issues.

K-Means

Thus far we’ve used machine learning m methods as tools to assist in the asset allocation process. In part one we used supervised learning processes to develop an indicator that provides us with a trading signal. Import to a we used unsupervised methods to help decompose our data into principal components thus allowing us to project volatility more easily and use less data in the process. But we have not seen is whether or not machine learning methods can actually allocate assets successfully on their own. To that end, model number three will use kmeans clustering on fundamental, valuation, and momentum data to generate a stock portfolio.

Features

Access

All the data required for this model was accessed via the refinish leave Eikon API and was stored in a series of CSV file for further use. Who is important to note that upwards of 250,00 data points were pulled for this model (505 S&P 500 Stocks, 21 features Per stock, over 24 quarters-pulling the data took over 45 minutes)

Features

As discussed previously we split our data into three subsections

Valuation:

In this section we pull the information for and calculate some of the most common stock market valuation multiples.

1. earnings yield- earnings yield is the reciprocal of the price to earnings multiple. This multiple is calculated by dividing a companies net income by their market cap, or using earnings per share and the stock price. For the sake of our calculation we chose to use forward earnings yield which uses analyst expectations for net income as opposed to historical data.
2. Book yield- this is calculated using the companies total book value in dividing it once again by there market cap
3. EBITDA to EV: EBITDA stands for earnings before interest taxes depreciation and amortization and EV stands for enterprise value. EBITDA is often used as a proxy for cash flow, well enterprise value is calculated by adding a companies market cap to its net debt. (Net debt equals total debt minus cash) to calculate our multiple simply divide EBITDA by EV
4. EBIT to EV: this is similar two the feature above, except that first multiple does not capture depreciation and amortization more closely resembles operating profits. The reason for including both is that companies that operate in industries with a high capital requirement that have significant depreciation expenses May unfairly advantaged via the EBITDA number, or unfairly disadvantaged via the EBIT number.

Fundamental

These features focus on company fundamentals, economic output, and cost of capital.

1. return on invested capital. Return on investment capital is often considered one of the most important indicators about the quality of a businesses. It is calculated as the net operating profit after taxes divided by A companies total invested capital.
2. Weighted average cost of capital: The weighted average cost of capital is an estimate of what a company has to pay it’s investors. It is in affect the financial cost of doing business for a company. The WACC is usually calculated by taking all the components of a company’s capital structure calculating their cost to the company, and calculating a weighted average of those two. There are in most cases considerations for taxes that must be made, and estimates for things like the cost of equity.
3. Long-term expected growth. This is nothing more than Wall Street’s aggregate estimate for expected growth for a given company over the next five years.
4. Total debt to total capital this is a measure of a company‘s financial leverage and is used to evaluate risk

Momentum

in this section we focus 36 and 12 months measures of momentum for a particular stock.

1. Risk adjusted relative momentum risk-adjusted relative momentum is calculated by subtracting a stocks total momentum over a given period of time from the benchmarks risk-adjusted return over the same timeframe. To adjust for risk, we multiply the stock’s beta coefficient (coefficient from a linear regression) to its benchmark.
2. Relative momentum Standard relative momentum is simply the returns of a stark ovary given period of time less the returns of the benchmark over the same time frame.

A good stock should have strong relative and strong relative risk-adjusted momentum, whereas a poor stock will likely not have strong risk-adjusted momentum

Feature prep

Winsorization- Financial data is often plagued with extreme observations. It is not uncommon to see price to earnings multiples in the thousands, stock momentum that has outpaced the rest of the market five or six fold, or extreme amounts of leverage on a companies balance sheet. The Windsor is ation process allows us to deal with these problems. Windsor is ation simply replaces The extreme top and bottom x percent Of all observations with the The value associated with the X percentile. In other words assume A range of numbers between zero and 100. If we were to Windsor rise the top and bottom 5%, then observations 95 three 100 would now equal 95, and observations five through would now equal five. We apply a 5% limit to the bottom and a 20% ceiling to the top in this model.

Main mix scaling -The mid mix scaling processes used to normalize features and minimize anomalies.

We evaluated three different potential clustering models for this segment. The first was TB skin, Second second agglomerative clustering, and the third Kmeans. I selected the kmeans. Do you skin produced for too many outliers given the nature of the Hi levels of dispersion in the data. Agglomerative clustering produced reasonable results but given the fact that the data in question does not really have a tree type structure Kmeans appeared to be the better choice.

Process

We set the queue means parameters to create four clusters. We fit the model using cross-sectional data for all stocks in the S&P 500 at a particular point in time.

Once the model has created the clusters, we group the stocks based on their clusters and purchase equal weights in the cluster that had the highest average score across all three feature categories value momentum and fundamental. We run the same model every quarter and rebalance the portfolio every time.

Results

A 3-D plot of the clusters shows that Kmeans is clearly identified for groups of Stocks. The challenge initially was deciding which group to pick as a portfolio. It is after all never clear which of the three factors we evaluated in this model or in fact driving the market. But further evaluation of the featured data relative to the clusters reveal that for the most part there is always one cluster that is slightly better than the rest.

Real world applications

The Caymans portfolio performed extremely well. Between 2016 in January 2021, The portfolio posted a gain of 117.49% versus 65.08% for the S&P 500 equal weight portfolio. From a risk-adjusted perspective, The sharp ratio over the period in question was 5.88 for the two means portfolio in 2.99 for the equal weight S&P 500 portfolio.