Stock Market Portfolio Optimization

Sahil Patel, Cole Reynolds, Neal Vaghasia

COGS 109, Professor Eran Mukamel

December 4th, 2018

Abstract

The potential ramifications of applying data analytics to the study of the stock market are enormous and could change the way people invest forever. In this project, we will use data analytics to construct an optimal asset portfolio consisting of stocks in the S&P 500 (the 500 largest companies in the United States). We do this through two different clustering techniques, Kernel Density Estimation and K-Means, to identify the top performing stocks based on their Sharpe Ratios (a metric that compares returns to risk). We determine the effectiveness of our two portfolios by comparing them to the index fund SPY (an approximation for the stock market as a whole). We find that the KDE portfolio is significantly the best, followed by the K-Means portfolio, then SPY. The KDE portfolio showed a 173% return, the K-Means portfolio returned 84%, and SPY returned 39%.

Introduction

Predicting the trends of the market is an essential task for both investors and businesses. The market is full of uncertainty and is affected by many unknown factors, so trying to predict its future performance can be a seemingly insurmountable task. The essence of this difficulty is captured by the generally accepted assumption that over long periods of time, the investor that actively tries to predict the future will perform worse than the investor that just lets their money sit in the market. In order to minimize risk while maximizing potential returns in this hazardous environment, many investors have turned to Modern Portfolio Theory (MPT) while constructing their stock portfolios. The basis of MPT is that the performance of a stock can be calculated in terms of its past returns and volatility (a proxy for risk). This calculation yields a number called the Sharpe Ratio, with higher values meaning more returns for less volatility. Using the framework of MPT, this project seeks to answer the following question: can investors

create an optimal stock portfolio that minimizes risk while maximizing returns and beating the performance of the general market?

We will seek to answer this question by applying data analysis techniques to framework of MPT. Specifically, we will use two clustering methods, Kernel Density Estimation and K-Means clustering, on the Sharpe Ratios of every stock on the S&P 500, then take the stocks from the topmost clusters and use them in our portfolio. We hope that our method will be useful to the average investor who does not have enough time nor the skills necessary to fully study the market and how it works.

Materials and Methods

In order to thoroughly test our methods, we used a dataset from Kaggle that covered all of the stocks on the S&P 500 from April 1, 2013 to December 30, 2016. This data was collected from the Investor's Exchange API, an open resource that people can use to pull accurate market data. Since this data was reported straight from the market and is not real-time, there is no noise present. Each observation consisted of the opening price, closing price, high price, low price, the trading volume, the stock's ticker symbol, and the day. The dimensions of this dataset are 851264 observations x 7 variables. In accordance with the assumptions of MPT, the only variables we used while calculating our predictor (Sharpe Ratio) were the closing price, ticker symbol, and date.

We analyzed the data using a novel rolling-window clustering technique. In this technique, we first segment the data temporally into 15 three-month quarters, then use each quarter as a training data set for the clustering models to determine the optimal stock portfolios for the next quarter. This effectively creates a succession of training and testing windows whereby we train on the first quarter, then test on the second quarter, then train on the second

quarter, then test on the third quarter, and so on. This is actually a modified version of cross validation, where we use each segment of the data for both training and testing.

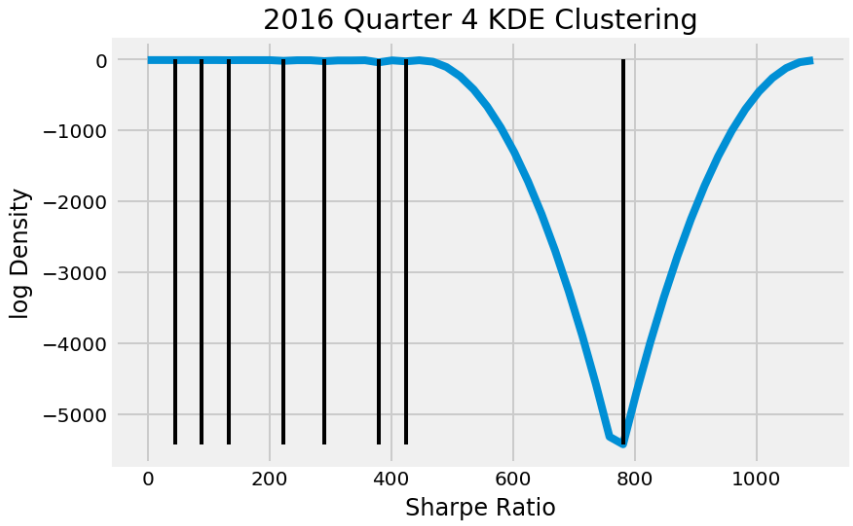
As we rolled through the windows, we used two models to generate different stock portfolios and tracked their individual performances. For the first model, we used Kernel Density Estimation (KDE), a clustering method that determines the distribution of the data and clusters it by splitting at the minima of the distribution. This model worked extremely well on our Sharpe ratio predictor because it is meant for 1-dimensional data. After repeated performance trials, we determined that any stock with a Sharpe Ratio greater than the fourth largest minima would be included in our portfolio. We used scikit-learn's KDE model for these calculations. For the second clustering model, we chose K-Means. K-Means was utilized in order to cluster the Sharpe Ratios of the stocks into five clusters. Once the data was clustered, we used all of the stocks in the cluster with the greatest Sharpe Ratios in our portfolio. Again, scikit-learn provided the model algorithm for these calculations. These two models were trained on a 1-dimensional vector of the Sharpe Ratios for each stock in a given quarter.

We started our test on April 1, 2013 with $100,000. After that, at the beginning of every quarter we would distribute our money by "buying" every stock that the two models said would perform well and calculate their respective returns. Then, at the end of the quarter we would "sell" all of the stocks and calculate the ending value of each portfolio.

This analysis technique was appropriate for the problem we had because we effectively utilized known past information to make educated guesses about future performance by using the framework of MPT. Because stock market data is limited in that it only consists of daily prices, it was necessary for us to derive a way to calculate the performance of each stock in a comparable way.

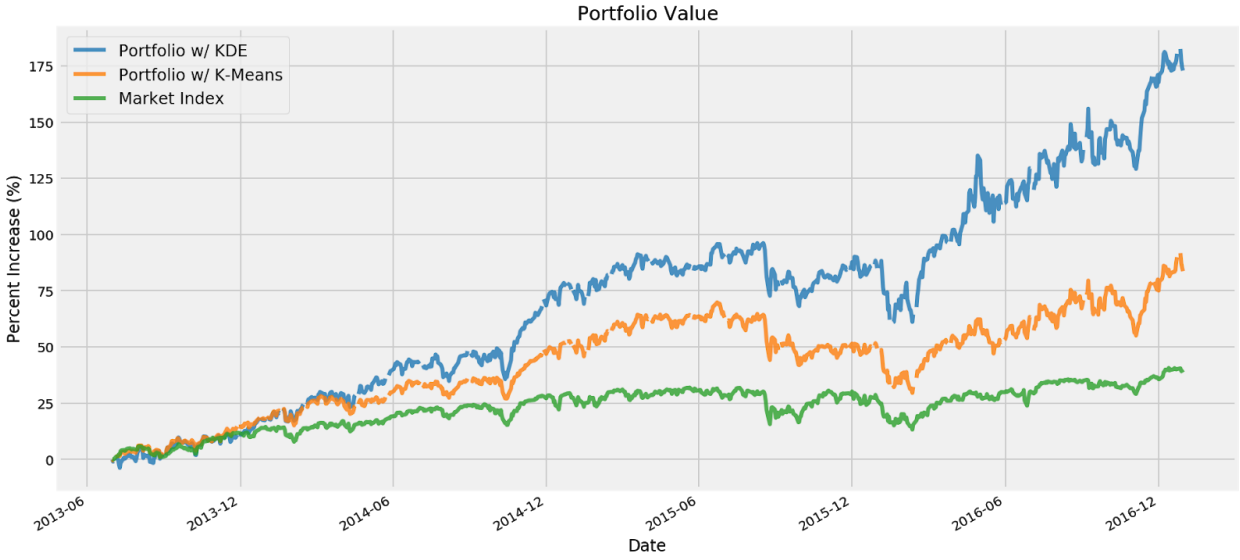
Results

Below is an example of the KDE clustering method's clustering decision. The vertical black lines denote the cluster separations and the blue line denotes the distribution of the data.



The KDE portfolio had an average of 10 stocks in the portfolio every quarter, while the K-Means portfolio had an average of 29 stocks in the portfolio every quarter.

Below is a plot of the performance of the portfolios and the market index SPY.



The KDE portfolio had a final return of 173% on initial investment, while the K-Means portfolio had a final return of 84%. Both of these numbers are significant, given that our benchmark SPY had a return of 39% over the same period.

With these returns, we find that it is possible to make an optimal stock portfolio that outperforms the general market while minimizing risk and maximizing return.

Discussions/ Conclusions

This project has the potential to make a significant impact in the ability for investors to make money in the market. The stock market can be a daunting thing, and this project and the technique outlined within it can significantly simplify the task of investing for everybody. Moving forward, we hope to see more people relying on data analytics to make rational choices about their investments instead of relying on blind faith or rumour.

Works Cited

Jagerson, John A. “Sharpe Ratio.” *Investopedia*, Investopedia, 24 Oct. 2018, www.investopedia.com/terms/s/sharperatio.asp.

Nugent, Cam. “S&P 500 Stock Data.” *RSNA Pneumonia Detection Challenge* | *Kaggle*, 10 Feb. 2018, www.kaggle.com/camnugent/sandp500?fbclid=IwAR2zstxrD09ceFhsYkoJfjcjasIYXm8nP3gV23w00_fPYRdJe7LdD1AAMw8.

“Sklearn.cluster.KMeans¶.” *1.4. Support Vector Machines - Scikit-Learn 0.19.2 Documentation*, scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

“Sklearn.neighbors.KernelDensity¶.” *1.4. Support Vector Machines - Scikit-Learn 0.19.2 Documentation*, scikit-learn.org/stable/modules/generated/sklearn.neighbors.KernelDensity.html#sklearn.neighbors.KernelDensity.