5-2022

Comparing Actively Managed Mutual Fund Categories to Index Funds using Linear Regression Forecasting and Portfolio Optimization

Luke Weiner
University of Arkansas, Fayetteville

Citation

Comparing Actively Managed Mutual Fund Categories to Index Funds using Linear Regression Forecasting and Portfolio Optimization

Luke Weiner

Undergraduate Honors Thesis in the Department of Industrial Engineering

University of Arkansas

Research Mentor and Thesis Advisor: Gregory S. Parnell, Ph.D.

Thesis Committee Members: Gregory S. Parnell, Ph.D., and Justin R. Chimka, Ph.D.

Comparing Actively Managed Mutual Fund Categories to Index Funds using Linear Regression and Portfolio Optimization

Abstract:

The global investment industry offers a wide variety of investment products especially for individual investors. One such product, index funds, which are younger than actively managed mutual funds, have typically outperformed managed funds. Despite this phenomenon, investors have displayed a tendency to continue investing in actively managed funds. Although only a small percentage of actively managed funds outperform index funds, the costs of actively managed funds are significantly higher. Also, managed fund performances are most often determined by their fund category such as *growth* or *real estate*. I wanted to answer the following question for individual investors: can we forecast the future performances of actively managed funds taken from multiple categories and build an optimized portfolio to outperform index funds. The goal of my research was to provide quality information to individual investors and to gain investment knowledge myself so that I can make wise investments in the future. Through my analyses, I discovered that creating fund forecasts often results in high error rates and requires macroeconomic factor stabilization, and global events can alter forecast accuracies severely. When optimizing a portfolio using returns, I determined that a constraint must be added to require diversification. Based on my results, individual investors should identify a broad spectrum of possible funds to invest in, select simple factors to make price predictions, be hesitant to respond eagerly to price forecasts, and understand how much return they are willing to give up for diversification. As an individual investor myself, this study gave me the knowledge to think far more strategically about my own investments and challenged me to understand my own risk tolerance.

Table of Contents

1. Background and Significance

The United States fund industry, comprised of financial products such as index funds, actively managed mutual funds, and electronically traded funds, was worth approximately $28 trillion in late 2020 . Within this industry, actively managed funds are on average five times more expensive to own than index funds [6]. However, index funds have historically outperformed most actively managed funds [3], and as of 2019, more assets have been invested in index funds than managed funds [9]. Several hypotheses were raised in the late nineties as to why consumers continue to invest in actively managed funds instead of index funds, which include their generally trusting view of brokers and their misunderstanding of what is acceptable fund performance [5]. Therefore, a comparison between the two is necessary to inform individual investors.

A study published in the Journal of Applied Business and Economics performed a comparison of two theoretical investment portfolios, one consisting of dividend-yielding quality stocks and one which was invested in the S&P 500 Index. Using dividend reinvestment, the quality stock portfolio produced a yield of 134.58% while the index portfolio resulted in a 69.59% yield. [4] Furthermore, dividend stocks often have decreased volatility, small price-to-book ratios [13], and better performance in "poor investment environment(s)" [2] such as recessions. Volatility is an indicator of investment risk [1], and the price-to-book ratio compares the total value of all company stocks to the total book value of the company [8]. Lower price-to-book ratios usually reveal that stocks are undervalued and have the potential to be "good investments" [8].

Although 52% of American families were invested in the market as of 2016, only 14% owned individual stocks. Most investments were also tied up in retirement accounts like a 401(k) [12], which most often only offer mutual funds [10]. Furthermore, many mutual funds exist that have a "high-dividend-yield," operating similarly to the dividend stocks they are comprised of. Therefore, a similar comparison could be made by substituting dividend stocks with high-dividend mutual funds.

Because a large amount of American families' money is tied up in retirement accounts dependent on the performance of actively managed funds, it is important to provide beneficial investment information to individual investors. Although a portfolio comprised of dividend quality stocks may outperform index funds, it is unhelpful to most common investors who shy away from individual stocks. However, it could be advantageous to American retirement accounts to compare index funds to actively managed mutual funds that have high proportions of quality stocks and high dividend yields and actively managed funds from different categories as well.

2. Outline of Research

The following process chart in **Figure 1** provides a high-level overview of the procedure I used for this research. Each level or step corresponds to the following major sections, respectively: *Selecting Funds Used for Analysis*, *Developing Forecasting Methods*, and *Optimizing Fund Portfolios.* As shown in the figure, I began by selecting six funds and six macroeconomic factors that I would use to forecast prices. Using these, I then created multiple price forecasts using linear regression techniques and selected a forecasting model for each of the six funds. Next, I used the forecasted returns in a portfolio optimization program that I created in Java. The output of this final step was a portfolio that maximized returns based on potential fund returns and a portfolio that maximized returns based on actual fund returns. Each of the major sections includes a detailed process chart corresponding to each of the three steps in Figure 1.

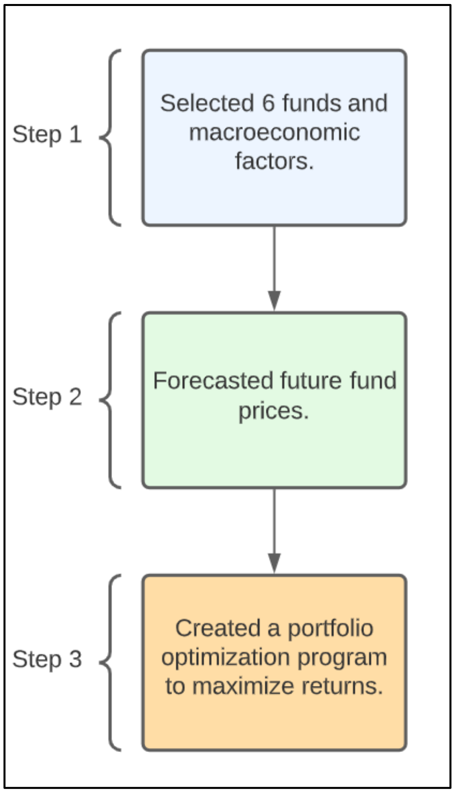

Step 1
Selected 6 funds and macroeconomic factors.
Step 2
Forecasted future fund prices.
Step 3
Created a portfolio optimization program to maximize returns.

Figure 1. *The high-level process chart for the three steps taken to complete this research project.*

3. Selection of Funds Used for Analysis

This section provides a summary of my decision process for selecting funds for my analysis. Although one fund selection was made using the weighted value model decision process, the majority were made based on observations of fund price history and graphical fund stability. **Figure 2** below provides a detailed process chart corresponding the selection of funds and macroeconomic factors as shown in step 1 of Figure 1.

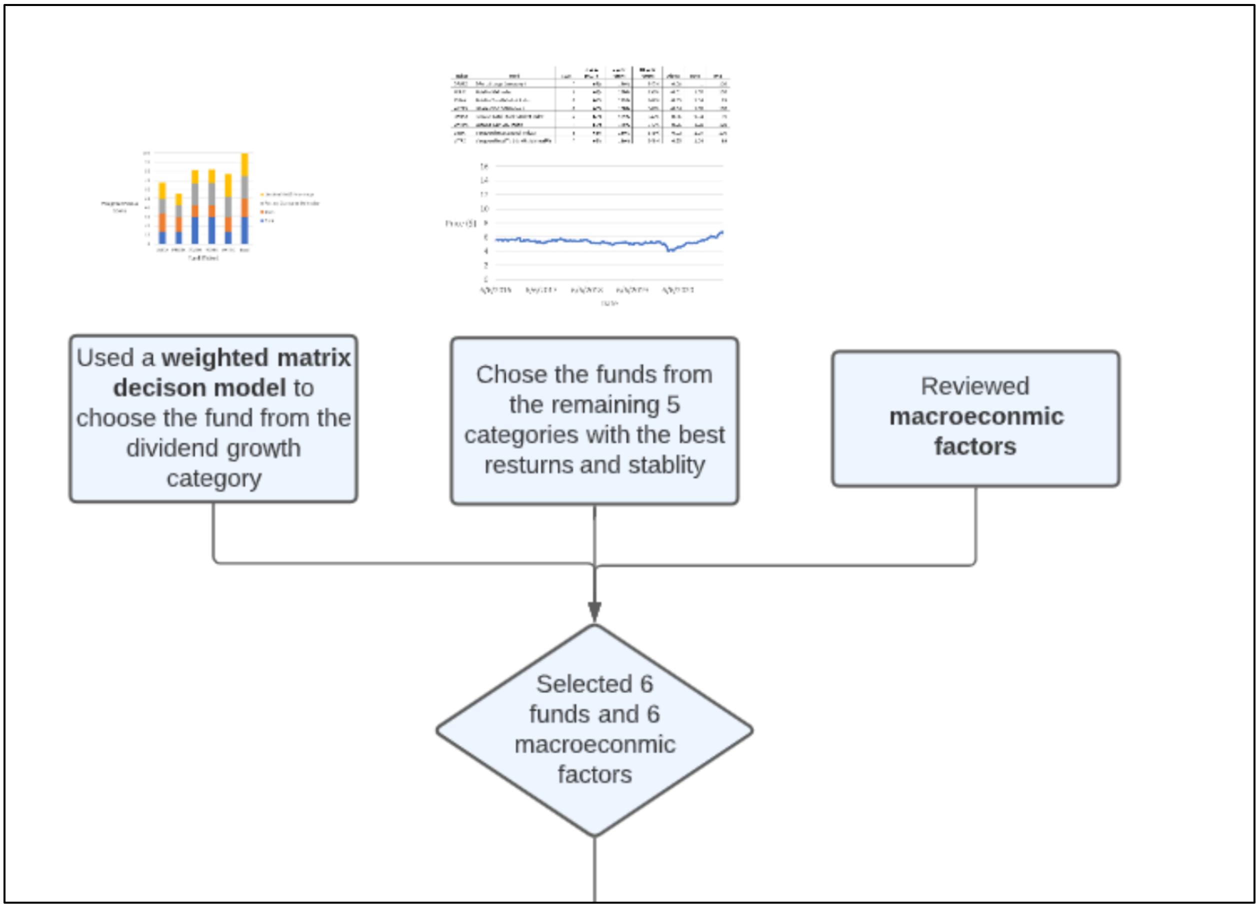

Used a weighted matrix decison model to choose the fund from the dividend growth category
Chose the funds from the remaining 5 categories with the best resturns and stablity
Reviewed macroeconmic factors
Selected 6 funds and 6 macroeconmic factors

Figure 2. *The process outline for selecting the six funds and macroeconomic factors used for analysis.*

3.1 Initial Fund Data Source and Decision Method

Initially, I sought to make comparisons of high-dividend funds and portfolios comprised of these funds. I began this process selecting seven funds from Morningstar's high-dividend yield list based upon historical rank of the last three years. All funds as shown below in **Table 1** were either of a silver or gold rank, which describes their future performance. Furthermore, the Star ranking is featured in Table 1 as well, which is a measure of past performance relative to the types of funds.

Table 1. *The Morningstar Rank and Stars for the chosen high-dividend yield funds.*

Fund	Ticker	Rank (Future Performance)	Stars (Past Performance)
BlackRock Equity Dividend	MADVX	Silver	3
ClearBridge Dividend Strategy	SOPYX	Silver	3
Columbia Dividend Income	LBSAX	Silver	5
T. Rowe Price Dividend Growth	PRDGX	Silver	4
Vanguard Dividend Appreciation Index	VDADX	Gold	3
Vanguard Dividend Growth	VDIGX	Gold	3

Vanguard High Dividend Yield Index	VHYAX	Silver	4

Next, I collected the following metrics for the funds shown in Table 1: percent consumer defensive, dividend yield percentage, and minimum investment, which are shown below in **Table 2**. I used the minimum investment of the funds as screening criteria. Funds were rejected if their minimum investment was over $10,000 to accommodate individual investors.

Table 2. *The percent consumer defensive, dividend yield percent, and minimum investment for the chosen high-dividend yield funds.*

Fund Ticker	**Percent Consumer Defensive**	**Dividend Yield Percent**	**Minimum Investment**
MADVX	8.70%	2.41%	$2,000,000
SOPYX	8.53%	2.28%	$1,000,000
LBSAX	10.05%	2.29%	$2,000
PRDGX	8.36%	1.63%	$2,500
VDADX	15.42%	1.85%	$3,000
VDIGX	15.80%	1.88%	$3,000
VHYAX	14.45%	3.11%	$3,000

Next, I normalized these factors for each fund. This was done to give the fund with the greatest value of a considered factor a score of 1.00 on a scale from 0.00 to 1.00. For all metrics except for Rank (gold or silver), this was calculated by dividing a funds value by the maximum of all fund values of each metric. To quantify the Rank, funds were given an initial value and then squared. I squared the values of the Rank metric because I did not consider the value curve for the Rank to be linear. I chose to use an increasing return-to-scale value curve where the initial values are squared because the incremental qualitative value between silver and gold seems to be greater among investors than the incremental value between bronze and silver. The initial values for the Rank metric were 1.00 for gold, 0.66 for silver, 0.33 for bronze (not used), and 0.00 for a neutral rating (not used). The values for the described fund normalized metrics are shown below in **Table 3**. Also, MADVX and SOPYX were removed from consideration because their minimum investments were greater than $10,000. Although, many large companies qualify for institutional investments for their employees [7]. Therefore, if an investor is considering applying this analysis to their company sponsored 401(k), they could keep higher minimum investment funds under consideration.

Table 3. *The quantified rank, star matric value, percent consumer defensive, and dividend yield percentage for the high-dividend yield funds under consideration.*

Fund Ticker	**Quantified Rank**	**Stars**	**Percent Consumer Defensive**	**Dividend Yield Percentage**
LBSAX	0.44	1.00	0.64	0.74
PRDGX	0.44	0.80	0.53	0.52
VDADX	1.00	0.60	0.98	0.59

VDIGX	1.00	0.60	1.00	0.60
VHYAX	0.44	0.80	0.91	1.00

After normalizing fund metrics values, I assigned weights for each metric as shown in **Table 4** below, and the sum of these weights is equal to 1.0. For each fund, a weighted score was assigned by multiplying the value of each of their normalized metrics by the specific metric weights as shown below in **Equation 1**.

Equation 1. *The weighted fund score equation used in the weighted decision matrix.*

$$If,$$

$$xi\ =\ normalized\ value\ of\ metric\ i\ for\ a\ specified\ fund$$

$$wi\ =\ weight\ of\ metric\ i$$

$$i\ ranges\ from\ 1, 2, \ldots, n$$

$$then, Weighted\ Fund\ Score\ =\ \sum_{i=1}^{n} w_i * x_i$$

Table 4. *The weight assigned to each of the metrics used in the weighted fund score.*

Metric	Weight
Rank	0.3
Stars	0.1
Percent Consumer Defensive	0.25
Dividend Yield Percent	0.25
Total	1.0

As shown in **Table 4,** Rank was given the largest weight because it is a reflection of the predicted future performance and to any investor this should be the most important aspect. However, the number of Morningstar stars logically seemed to hold a fraction of the importance of the Morningstar Rank because it describes past performance, which does not necessarily determine the future performance, and the range is only 3 to 5 stars. Furthermore, the percentage of funds comprised of consumer defensive stocks and the dividend yield percent were goals of using a high dividend yield mutual fund, but I never differentiated their importance in that goal, so I decided to equally weight them in the decision process. When considering the number of Morningstar stars a fund has, I chose not to set it as a screening criterion. If an investor constrains the funds under their consideration to only those with five stars, they are only constraining their choices to funds with superior past performance. However, this may forego the opportunity to invest in funds that are predicted to perform well despite undesirable past performance.

The resulting weighted scores for each fund are illustrated below in **Figure 3**, which displays the weighted score of each metric within each fund. Also, the ideal is also illustrated as a comparison. Because VDIGX, or Vanguard Dividend Growth Fund, had the highest weighted value score of about eighty-two, I chose it to be included in a further analysis for forecasting and portfolio optimization. Furthermore, a weighted value score of eighty-four can be interpreted as possessing eighty-two percent of the difference in value from the ideal fund hypothetical fund with and the lowest possible scores on each metric.

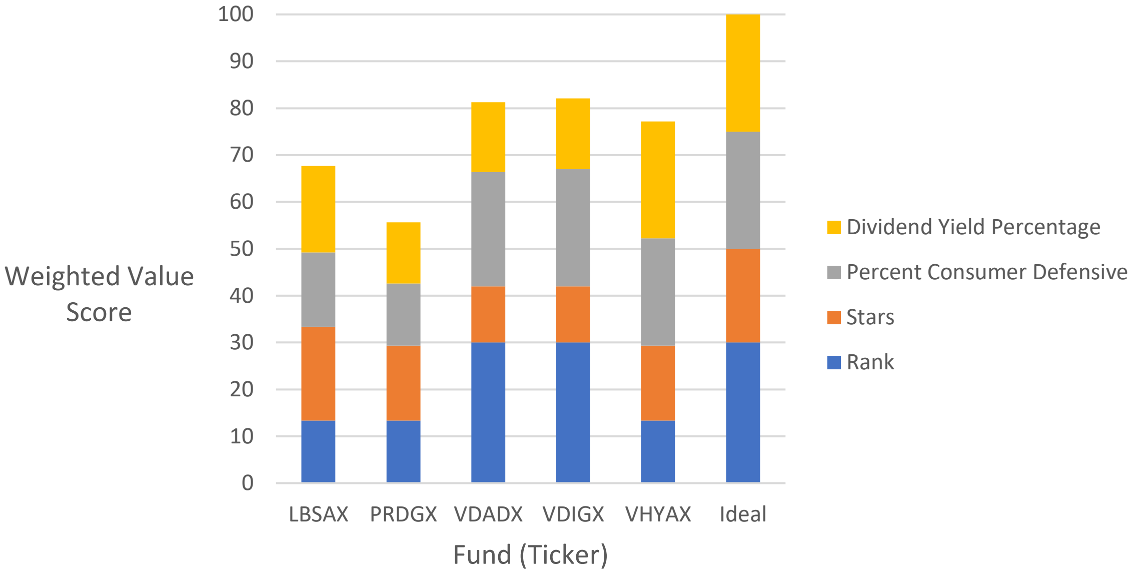


Figure 3. *The graphical representation of the resulting weighted value scores of each of the high-dividend yield funds under consideration.*

After completing this process, I decided that a comparison of multiple fund categories would be more beneficial than just using high-dividend yield categories. However, because the weighted decision method is more tedious in terms of data collection, I decided to utilize a more qualitative approach for determining the fund to use within the remaining categories. I also decided to use the following fund categories: commodities, growth, international large growth, high dividend yield, and index funds as the comparison.

3.2 Qualitative Decision Method for Fund Selection

Although I initially used a weighted decision process for selecting the high-dividend yield fund, I used a less formal approach to selecting the remaining five funds. For example, to select an index fund, I chose the top performing index funds from Morningstar as shown in **Table 5** below. Instead of using the previous factors as comparative metrics, I chose to use the Morningstar star rating, the three-year return, the five-year return, the ten-year return, the alpha and beta values, which were described in the background, and the R-squared value. To clarify, the alpha-value of a fund is a measure of its performance relative to an expected performance predicted by the beta-value, and the beta-value of a fund is a measure of its risk relative to the market, and the [11].

Table 5. *The metric values for 8 index funds used for the selection of an index fund.*

Ticker	Fund	Stars	3-year return	5-year return	10-year return	Alpha	Beta	R^2
DFUSX	DFA US Large Company I	4	63%	129%	347%	-0.08	1	100
FXAIX	Fidelity 500 Index	5	63%	130%	350%	-0.01	1.00	100
FSKAX	Fidelity Total Market Index	4	63%	129%	344%	-0.65	1.04	99
WFSPX	iShares S&P 500 Index K	4	63%	130%	348%	-0.04	1.00	100
SWTSX	Schwab Total Stock Market Index	4	63%	129%	342%	-0.66	1.04	99
SWPPX	Schwab S&P 500 Index	4	63%	129%	347%	-0.05	1.00	100
VINIX	Vanguard Institutional Index I	5	63%	130%	349%	-0.02	1.00	100
VITPX	Vanguard Instl Ttl Stk Mkt Idx InstlPls	4	63%	130%	348%	-0.55	1.04	99

By examining the table above, it was clear that two funds were the major contenders for being chosen: Fidelity 500 Index (FXAIX) and Vanguard Institutional Index I (VINIX). Both funds were rated the highest in terms of a star rating and had effectively equal 3-year returns, 5-year returns, beta values, and R-squared values. Because FXAIX had a higher 10-year return and alpha level than VINIX, I chose to use FXAIX for my forecasting and portfolio optimization. Although FXAIX outperformed VINIX, the difference was marginal, and VINIX would have served as a reasonable control fund as well as the other index funds. Because index funds track the stock market, performances differences are less due to stock picking and more due to the speed at which stocks are bought and sold to represent the market.

After an index fund was identified to be used as a comparison, funds from the remaining categories were needed. To select the remaining funds, I used the same premise of identifying top color ranked (gold, silver, bronze) funds by Morningstar and viewing their three, five, and ten-year performances. For most funds, my criteria were generally the ranking and performance, but I used the minimum initial investment as a screening criterion. Generally, if the minimum initial investment was greater than 10,000, the fund would not be considered because the goal of this research is to aid the individual investor. For example, when selecting a fund for the growth category, Harbor Capital Appreciation Fund Institutional Class (HACAX) would have been the prime candidate with a 10-year return of 212.5%, but its minimum initial investment is $50,000. Therefore, it would not be a feasible investment for most individual investors.

In process of selecting the remaining funds, I also looked at their price graphs to understand their stability as well as overall performance. This method proved useful specifically for the commodities funds, which are higher risk. For example, when comparing a bronze ranked fund, Parametric Commodity Strategy Fund Institutional Class (EIPCX) to a silver fund, PIMCO CommoditiesPLUS® Strategy Fund Institutional Class (PCLIX), I looked at their price graphs as shown in **Figures 4** and **5**, respectively. At the time that I observed these, which was in October 2021, EIPCX clearly had more stability, as shown by its graph. Changes in the price of EIPCX were not as large as the changes in price of PCLIX, especially on a shorter timeline. Stability was the largest criterion for commodities funds because commodities stock trading often have more inherent risks involved.

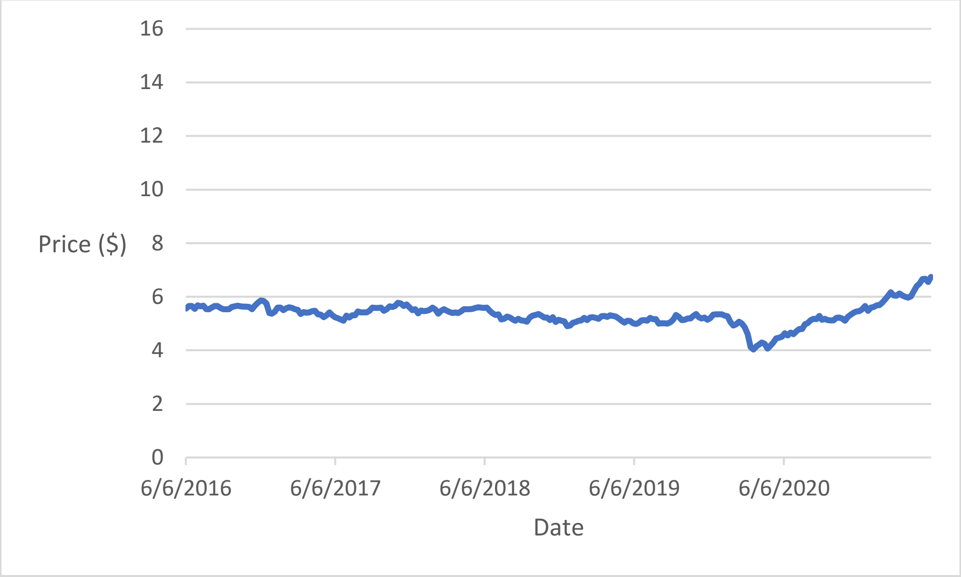


Figure 4. *The five-year weekly price plot for EIPCX (Yahoo Finance).*

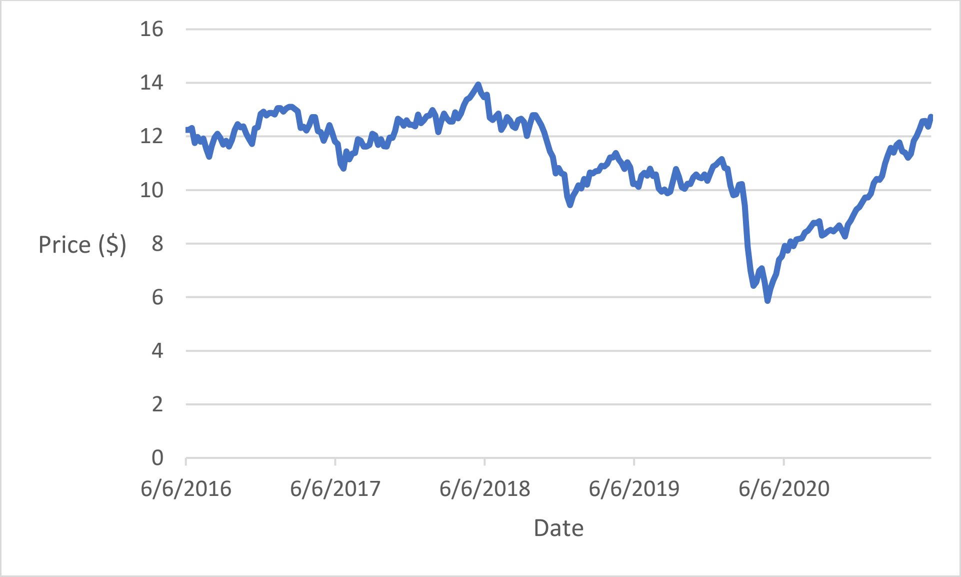


Figure 5. *The five-year weekly price plot for PCLIX (Yahoo Finance).*

As a result of the above processes, the funds shown in **Table 6** below were selected. Also, the additional data used for fund analysis was obtained from Yahoo Finance after they were selected.

Table 6. *The summary of the funds selected from each of the six categories.*

Category	**Fund**	**Ticker**
Dividend	Vanguard Dividend Growth Fund	VDIGX

International Large Growth	Hartford Schroders International Stock I	SCIEX
Growth	PRIMECAP Odyssey Growth Fund	POGRX
Index	Fidelity 500 Index Fund	FXAIA
Commodities	Parametric Commodities Strategy Fund	EICPX
Real Estate	Baron Real Estate Fund	BREFX

3.3 Macroeconomic Factor Data

To create forecasting models for the selected funds, I selected several macroeconomic factors to determine their effectiveness for predicting future fund prices. The macroeconomic factors were the average federal interest rate and the federal debt, taken from fiscaldata.treausury.gov, as well as the U.S. import/export price index, the U.S. labor productivity, the federal unemployment rate, and the urban consumer price index (CPI), which were taken from bls.gov. All macroeconomic factor data was standardized to fit a weekly projection. If the original dataset was set as a monthly basis, then each week within that month would take the value of that month.

4. Forecasting Methods

This section describes the methods I used to forecast future fund prices. Although no method used produced perfect results, the forecasts served as a basis for a later portfolio optimization.

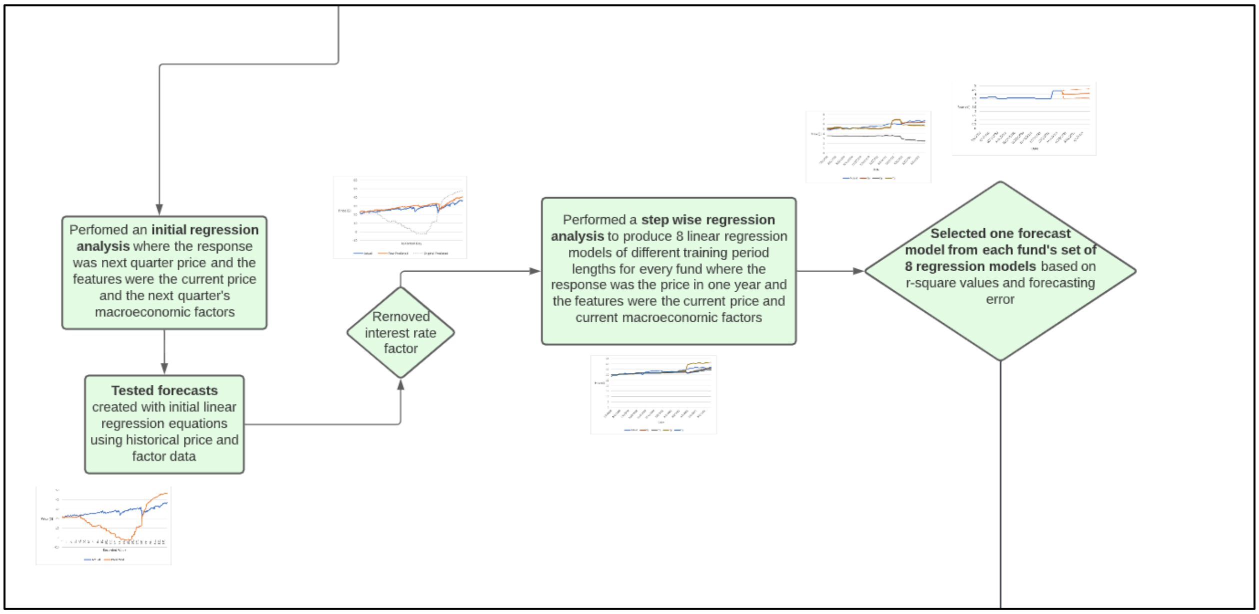

Perfomed an **initial regression analysis** where the response was next quarter price and the features were the current price and the next quarter's macroeconomic factors
Tested forecasts created with initial linear regression equations using historical price and factor data
Removed interest rate factor
Performed a **step wise regression analysis** to produce 8 linear regression models of different training period lengths for every fund where the response was the price in one year and the features were the current price and current macroeconomic factors
Selected one forecast model from each fund's set of 8 regression models based on r-square values and forecasting error

Figure 6. *The process outline for forecasting fund prices and selecting a forecast model for each.*

4.1 Using Current Prices and Future Macroeconomic factors

Originally, I attempted to forecast fund prices one financial quarter in advance using the current price as well as the future macroeconomic factors. I used Minitab to create linear regression equations for VDIGX, SCIEX, and POGRX using initial training data. The forecasts created for VDIGX, SCIEX, and POGRX using their regression models are shown below Figures 7, 8, and 9. The prices forecasted are for January 2016 through the end of June 2021, and I used approximately the previous five years of factor and funds data for my training set.

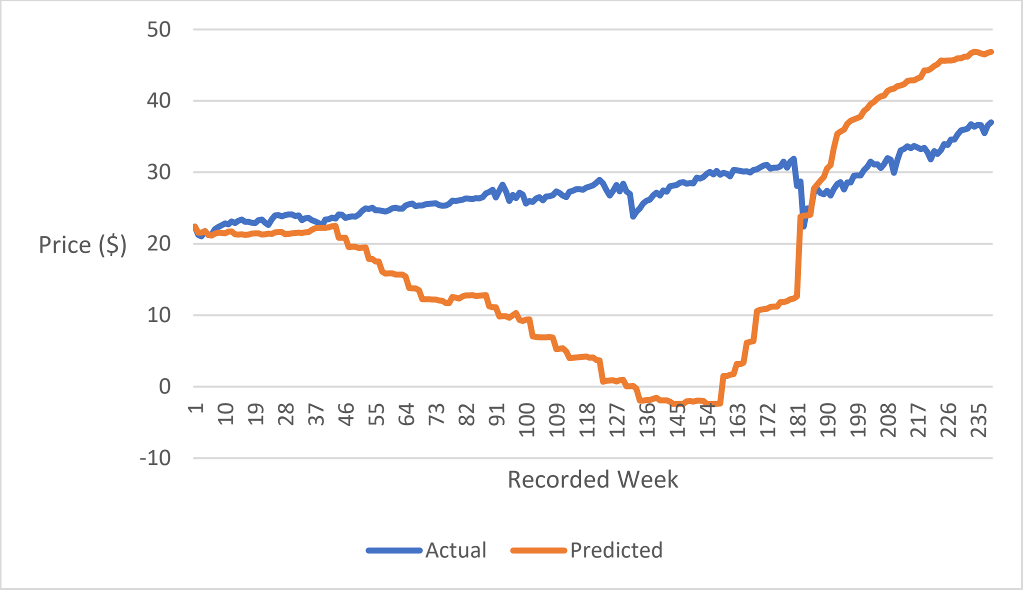


Figure 7. The forecast of VDIGX using the initial model with all macroeconomic factors.

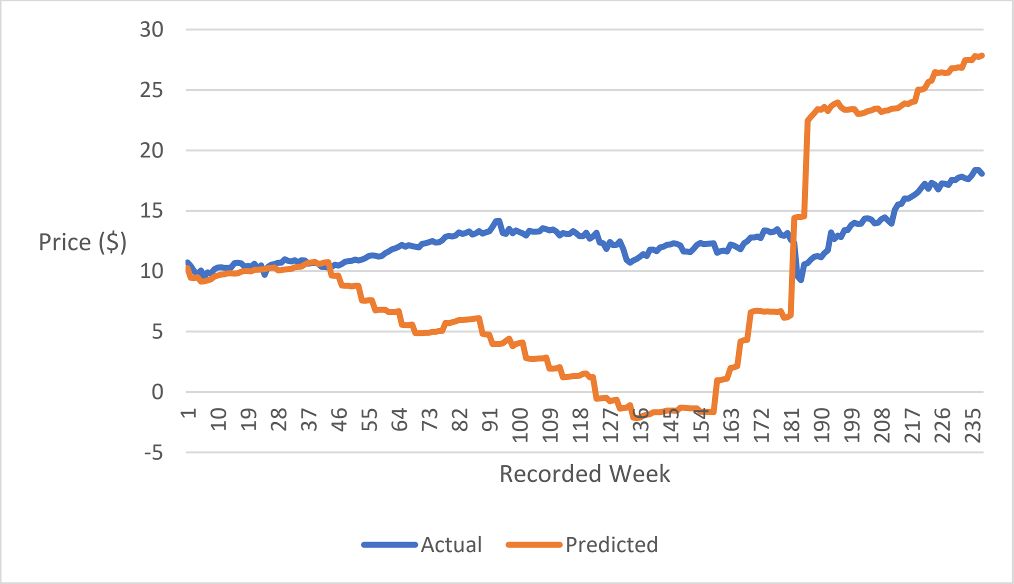


Figure 8. The forecast of SCIEX using the initial model with all macroeconomic factors.

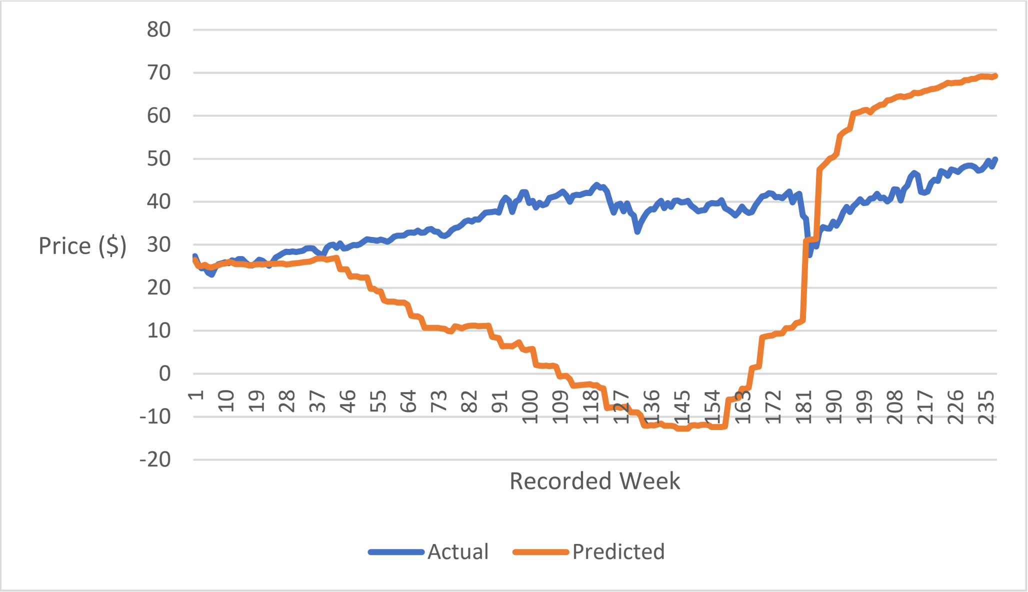


Figure 9. The forecast of POGRX using the initial model with all macroeconomic factors.

As shown in Figures 7, 8, and 9, the forecasts of each fund were dramatically skewed, but each one appeared to be skewed at the same point or recorded day and all readjusted around the same point as well. This led me to believe that a macroeconomic factor may have caused such error as it was shared by all funds. However, each fund may not have had similar reactions to the macroeconomic factor. After examining the data, particularly at the points in which the predicted prices fell and rose dramatically, it appeared that the interest rate factor increased 2 points and then decreased 2 points. This change is shown in **Figure 10** below. Note that the stable value level is differentiated from the dramatic fluctuation by color and the line in red is also the period over which my original forecasts were performed.

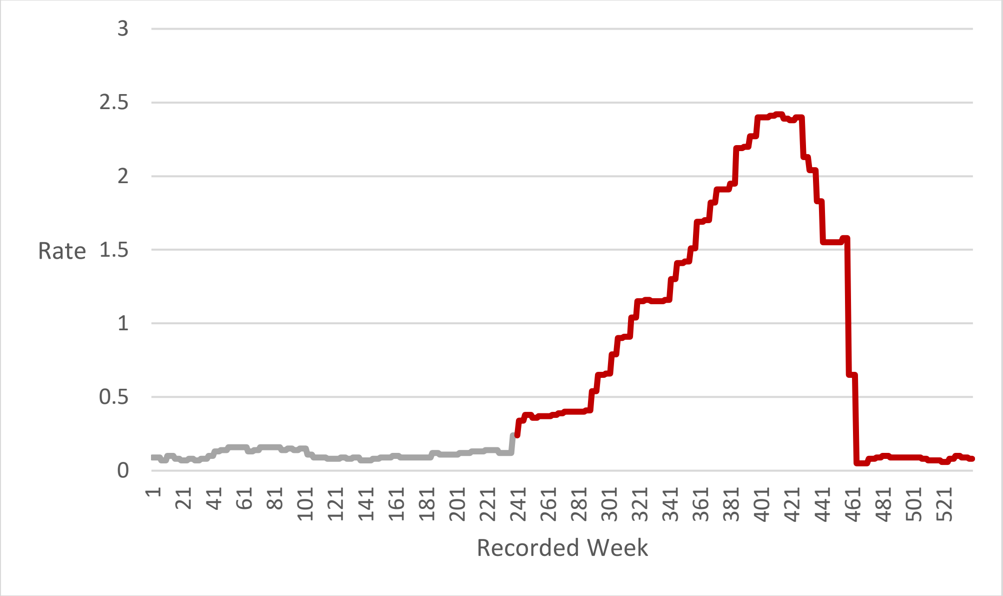


Figure 10. The federal interest rate plot from the forecasting period of VDIGX, SCIEX, and POGRX.

After determining that the interest rate experienced a 2% change during my forecasting period, I wanted to see the impact of its removal. For VDIGX, I performed the same forecast but did not include the interest rate. This is shown below in **Figure 8**, which compares the actual, previous forecast, and new forecast. Based on the results shown in the graph in Figure 8, I concluded that the forecast accuracy increased through the removal of the interest rate. Furthermore, because of this finding, I decided to remove the interest rate factor from future forecasting considerations.

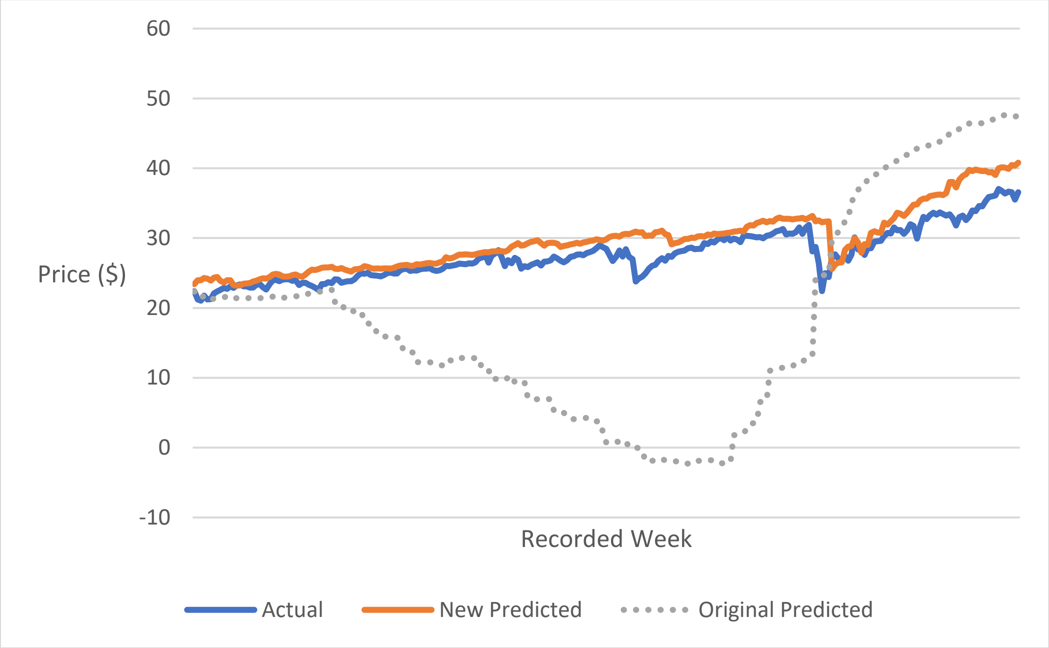


Figure 11. The comparison of the original forecast of VDIGX using all regression factors and the new forecasting with the interest rate removed.

Although this initial method produced a forecast that resulted in a low error, specifically a mean absolute percentage error of 7.6%, we identified two problems with it. First, although I removed the interest rate, I failed to also consider the possibility that not all macroeconomic factors should be considered as significant for each fund. To solve this problem, I used a stepwise model fitting procedure, which Minitab has the capability of doing automatically.

The other major problem associated with this initial approach was that I was forecasting future prices with the macroeconomic factor values of the forecasted period. Although this produced acceptably accurate results, in practice, this would require us to either obtain accurate predictions of those factors or create our own predictions. This added level of complexity would be impractical for future use and for the use of individual investors. Therefore, I chose to forecast the fund prices using the current macroeconomic factors. While this may not produce as accurate of results, it is far more practical in terms of using available information. Also, I altered my forecasting procedure to forecast one year in advance instead of one quarter because many individual investors reevaluate their portfolios annually.

4.2 Using a Stepwise Linear Regression Model

To begin my process for performing the stepwise linear regression model fitting, I first aggregated the fund and macroeconomic factor data into a standardized timeframe. This dataset included nine years of fund and factor weekly data, where the years used for training the linear regression model began at the beginning of July each year. The time frame of this dataset began in July of 2012 and ended in July of 2021.

To train the linear regression models for each fund, I used Minitab's stepwise regression function and decided to test eight different training sets for forecasting. While the forecasting and testing year was kept constant (July 2020 - July 2021), each of the eight training sets would create linear regression models using the associated number of years prior to the testing time frame. For example, the sixth training set would use the six years of data before July 2020 to build the regression model whereas the third training set would use the previous three years of data instead.

For each fund, my process for choosing the training set time frame was based on a combination of comparing model R-square values and assessing the accuracy of the forecasts through visual inspection and comparison of forecasting error. I chose to use the mean absolute percentage error as the forecasting error.

Shown in **Table 7** below are the linear regression equation coefficients obtained from each of the 8 training periods for the Vanguard Dividend Growth Fund (VDIGX). The R-square values are shown as well, which was particularly useful in determining which training period to use for further analysis. Because the longer training periods produced significantly greater R-square values, for this fund, I chose to plot the predicted prices of eight, seven, six, and five- year training periods, which is shown in **Figure 12**. For reference, predicted prices will be labeled with their training period length in years followed by a "p" (i.e., the eight-year training period is labeled as "8p").

Table 7. *The regression equation constants and R² values for the 8 step wise regression models created for VDIGX.*

Training Period (Years)	Equation Constant	Current Price Coefficient	Unemployment Rate Coefficient	CPI Coefficient	Debt Coefficient	Performance Coefficient	Exports Coefficient	R^2
1	221.8	0	0	-0.760	0	0	0	45%
2	14.34	-0.329	0	0	1.133	-0.491	0	29%
3	0.95	0	0	0	1.324	-0.3489	0	41%
4	-6.23	0	0	0	1.648	-0.1763	0	65%
5	-8.02	0	0	0	1.7336	-0.1785	0	78%
6	-14.22	0	0.617	0	1.901	-0.0986	0	80%
7	-17.55	0	0	0.0794	1.236	-0.0902	0	86%
8	-12.32	0	0	0	1.7425	-0.1226	0.0333	91%

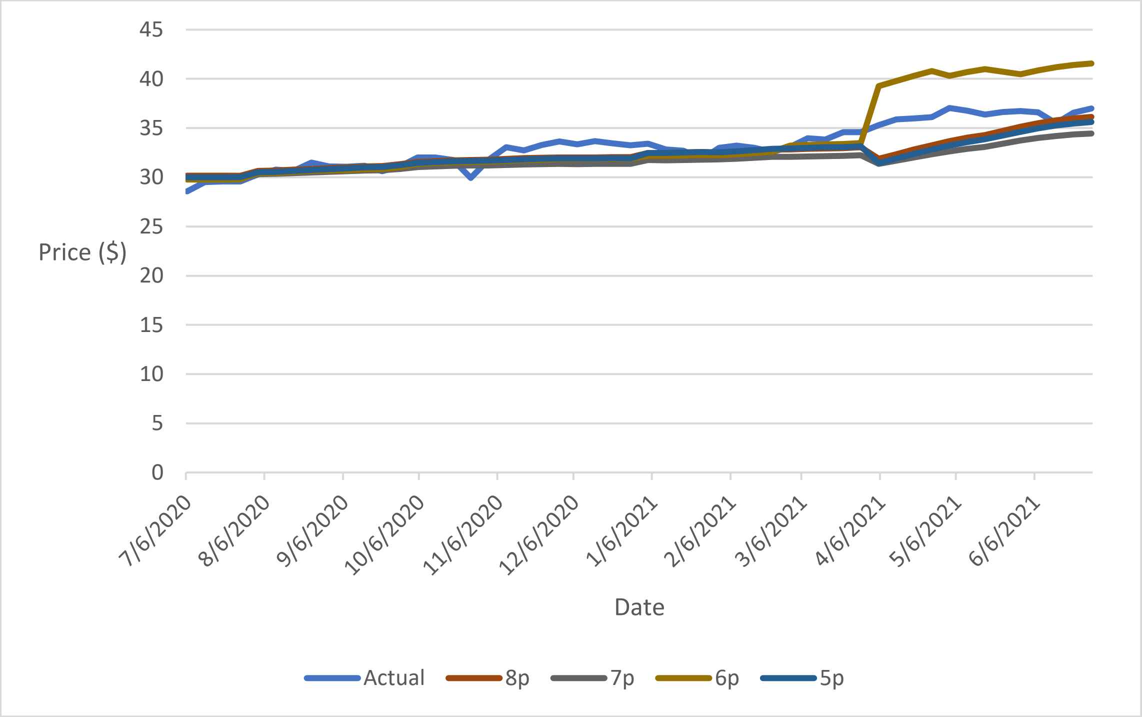


Figure 12. *The forecasted prices for the 8, 7, 6, and 5-year training period models of VDIGX.*

Based on the forecasting results in Figure 9, the 8p, 7p, and 5p forecasts were observably similar in terms of forecasting accuracy. Therefore, I differentiated them by their MAPE's. The MAPE's of the 8p, 7p, and 5p forecasts were 3.27%, 4.86%, and 3.56%, respectively. Because the forecasting error of the eight-period training set the lowest out of the three, I chose to use it for the portfolio optimization phase of this research.

Shown in **Table 10** are the linear regression equation coefficients obtained from each of the 8 training periods for the Parametric Commodity Strategy Fund Institutional Class (EICPX) as well as their respective R-square values. Unlike the other funds forecasted so far, the R-square values did not have as dramatic of an increase as the training period increased. Furthermore, as shown in the “R²” column, it was difficult to assess if a trend existed as the training period increased. Because of this, I took the five forecasting equations with the highest R-square values and compared their forecasts with the actual prices of EICPX. I compared more of the training period forecasts because the trend did not exist in the R-square values. For the previous funds, the increasing trends allowed for easier determination of the “best” fitting model. However, because most of the R-square values were close for EICPX, it required comparing more of the training period forecasts against the actual prices as shown in Figure 12.

Table 8. *The regression equation constants and R² values for the 8 step wise regression models created for EICPX.*

Training Period (Years)	Equation Constant	Current Price Coefficient	Unemployment Rate Coefficient	CPI Coefficient	Debt Coefficient	Performance Coefficient	Exports Coefficient	R²
1	59.46	0	0	-0.0678	-0.413	-0.1303	-0.2217	84%
2	22.70	-0.252	0	-0.06399	0	-0.1406	0	74%
3	21.20	0	0	-0.0758	0.1444	-0.1171	0	79%
4	23.68	0.2627	-0.323	-0.0875	0.1575	-0.0856	0	72%
5	7.79	0.2204	0	0.0264	-0.1645	-0.0570	-0.0551	49%
6	-22.70	-0.2973	1.3983	0.1211	-0.1650	0	-0.0405	76%
7	-24.86	-0.3501	1.5613	0.1344	-0.1541	-0.0214	-0.0383	89%
8	-20.39	-0.2290	1.5491	0.10512	0	0	-0.0468	93%

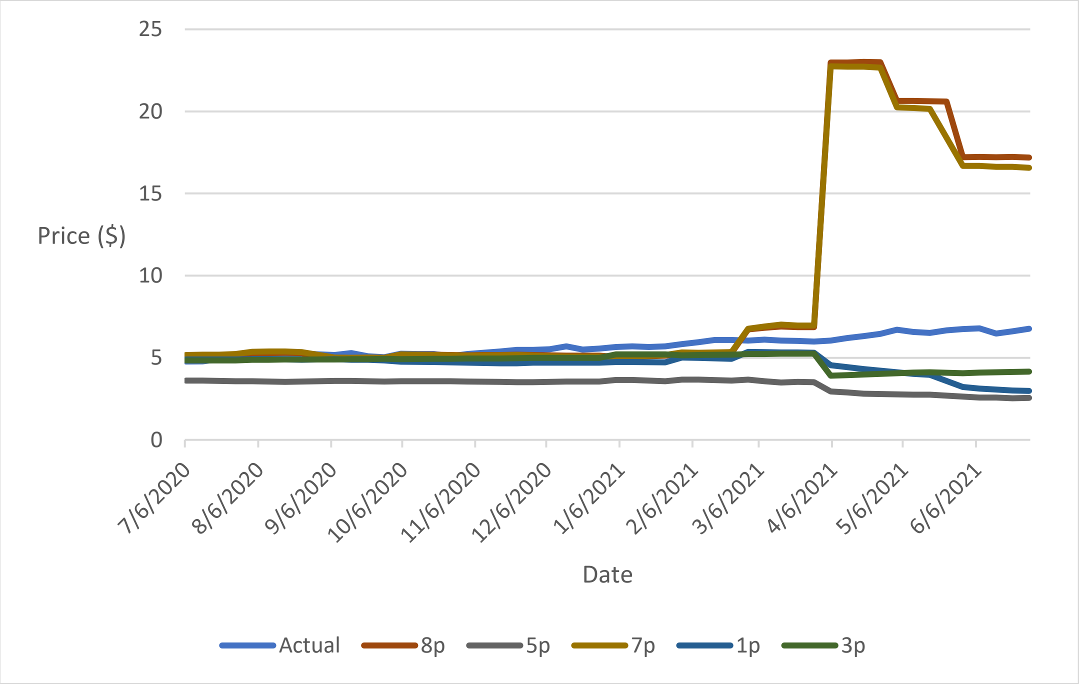


Figure 13. *The forecasted prices for the 8, 7, 5, 3, and 1-year training period models of EICPX.*

After calculating the forecasted prices for training periods included in Figure 13, it became difficult to choose a forecast. Qualitatively, I desired that the chosen forecast have an increasing trend because it would be used in my portfolio optimization, and there would be no purpose to including an investment that is forecasted to lose money. As shown, the 5p, 3p, and 1p forecast did not have an increasing trend, but their forecasting accuracy was higher compared to the 8p and 7p forecasts. Although the 8p and 7p forecasts had increasing trends, their forecasts predicted a dramatic shift in price, which is unlikely when considering mutual funds. To determine the reason behind this dramatic increase, I assessed the factors used by the 8p and 7p forecasts and determined if any had a dramatic shift. I determined that out of the five forecasting equations considered, only 8p and 7p were dependent on the unemployment rate, which experienced a dramatic increase during the beginning of April of 2021 as shown in **Figure 14**.

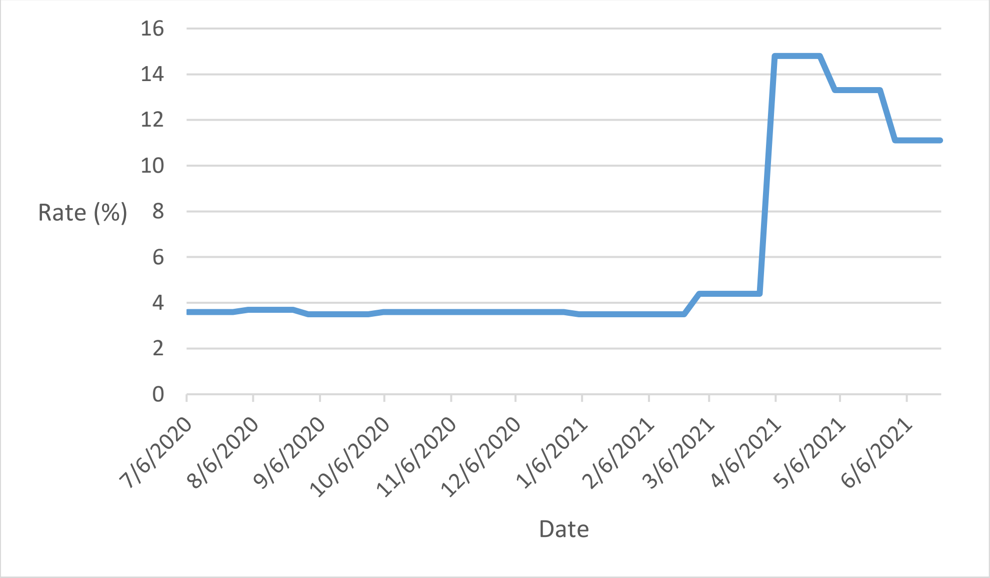


Figure 14. *The time series plot of the unemployment rate factor for the time period used in testing the forecasting models.*

This dramatic change in the unemployment rate represents forecasting error caused by uncontrollable events. This also represents the limitations of forecasting funds using linear regression. For my method of linear regression to have stabilized the forecast of EICPX with this dramatic increase in unemployment rate, the training data would have had to include far higher historical unemployment rates. However, for many modern mutual funds, they may not be old enough to have existed during periods of high unemployment, and economic reactions especially with investment vehicles like mutual funds, evolve in their reactions to macroeconomic factors. I do not believe funds two decades ago would have responded similarly to funds to today due to the ever-changing climate of our economy. Because I was unable to create an accurate forecast with positive returns, I chose to fit a more accurate model by changing the input values of the unemployment rate after the point in which they dramatically increased. As shown in **Figure 15**, I used Microsoft Excel's forecasting tool to forecast the unemployment rate values after they destabilized.

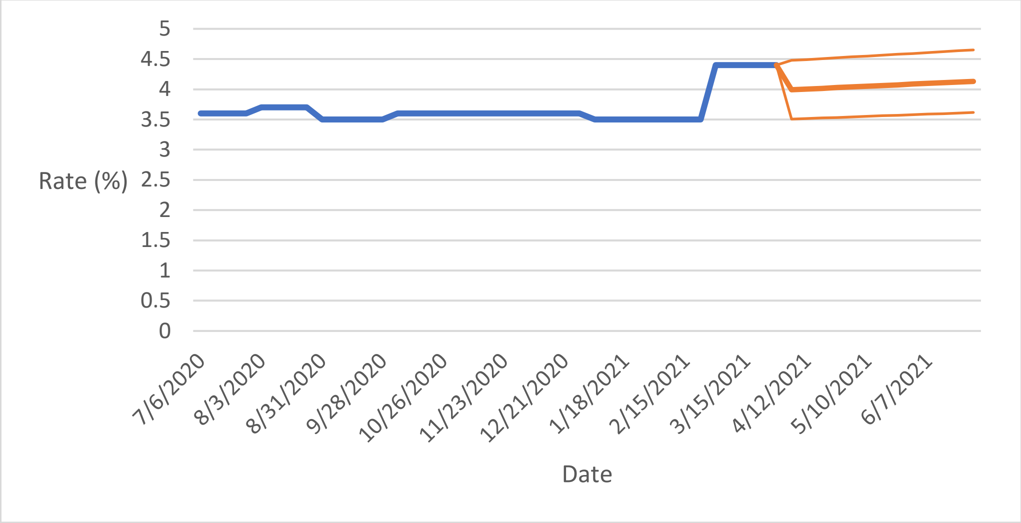


Figure 15. *The plot of the stabilized unemployment rate factor using Excel's forecasting tool.*

Once the unemployment rate values were changed, I plotted the 8p and 7p forecasts as well as the actual EICPX prices and 5p forecast as reference as shown in **Figure 16**. This model alternation produced a higher observable forecasting accuracy for 7p and 8p, which both now have a higher accuracy than the 5p. Also, as shown, the 8p forecast produced by this new model is the most accurate forecasting choice for my portfolio optimization. However, I needed to justify using an altered model in a portfolio because I had the benefit of knowing the actual prices and the need to change the model. I did end up choosing to use this altered model and the 8p forecast because the forecasts are based on current macroeconomic factors, and if current factors were to dramatically increase, I can now alter future predictions by discarding the dramatic changes and using more stable factor forecasts.

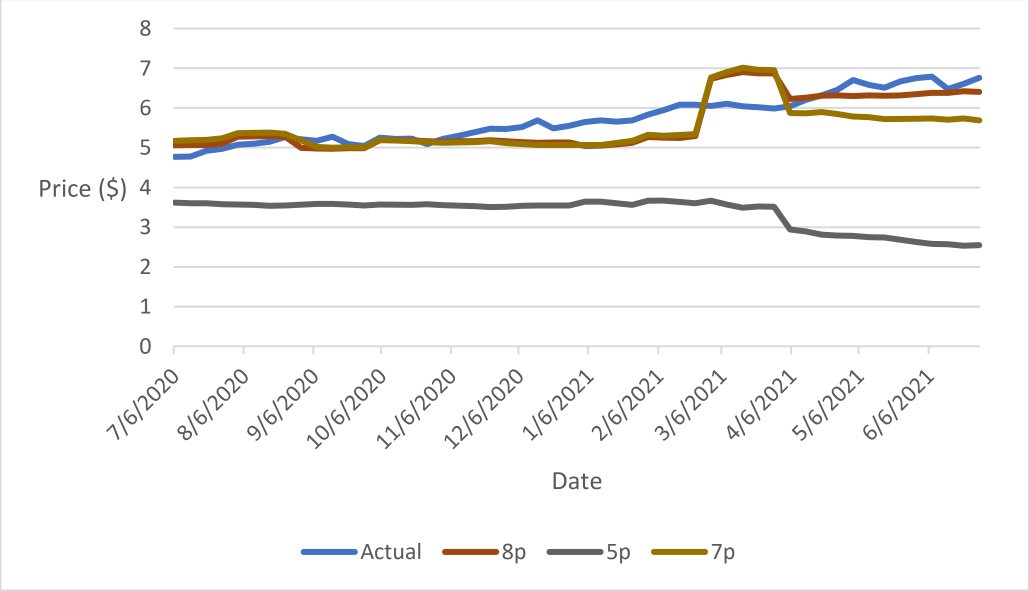


Figure 16. *The forecasted prices for the 8, 7, and 5-year training period models of EICPX using altered unemployment rate values.*

To summarize, I have included **Table 13** below, which include the model chosen for each fund to be used in my portfolio optimization. The detailed selection summaries for the remaining funds are included in the *Appendix* section.

Table 9. *The summary of the models selected for each of the six funds.*

Fund Ticker	Model Chosen	Predicted Return	Actual Return
VDIGX	8p	26.3%	29.3%
SCIEX	1p (first altered model)	21.7%	42.0%
POGRX	8p	23.9%	27.9%
FXAIX	8p	25.8%	38.5%
EICPX	8p (altered model)	36.3%	43.8%
BREFX	3p (altered model)	8%	49%

5. Applying Optimization to Fund Portfolios

This section describes the process I used to create optimized portfolios of the six funds chosen and their regression models. As shown in **Figure 17**, I first used the predicted 1-year return (Table 13) of each fund to create a portfolio with a maximum potential return within an investor's risk level or beta value. I then used that portfolio and calculated the returns based on the actual return of each fund (Table 13). Next, I created a portfolio that maximizes actual return instead of predicted returns and compared it with the actual returns of the portfolio constructed out of predicted returns.

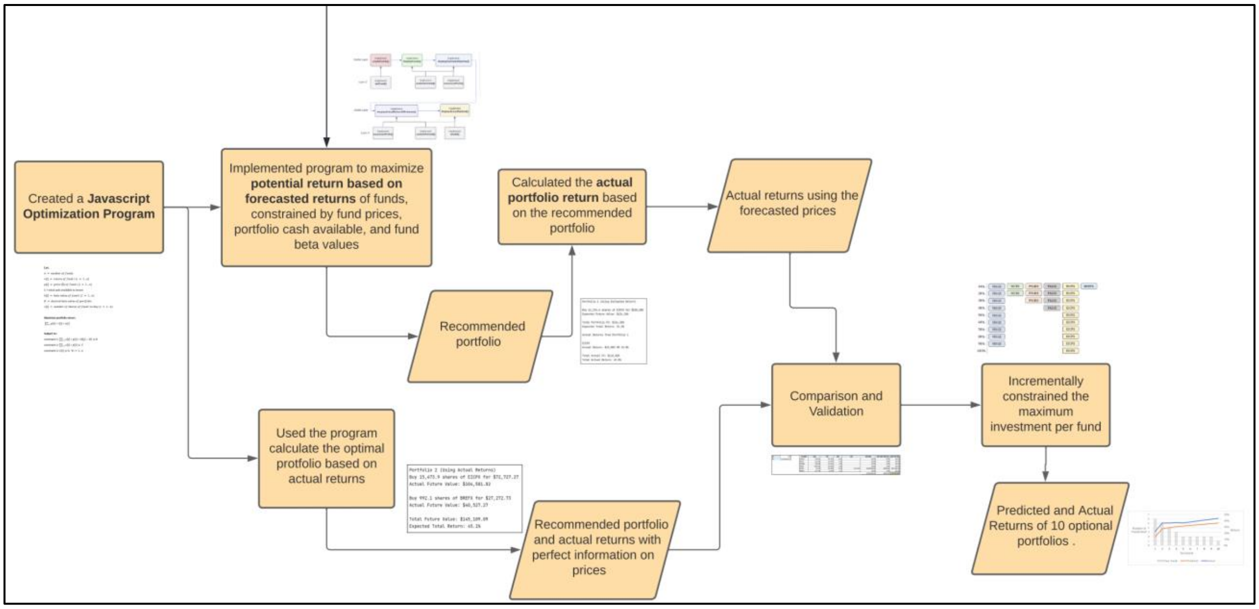
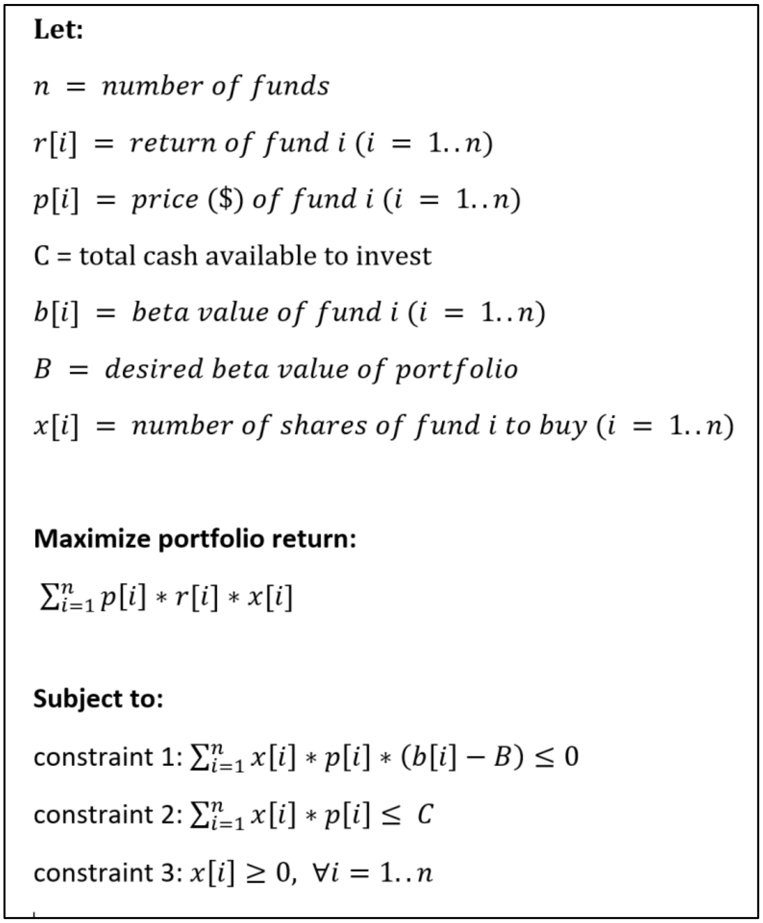

Created a **Javascript Optimization Program**
Implemented program to maximize **potential return based on forecasted returns** of funds, constrained by fund prices, portfolio cash available, and fund beta values
Calculated the **actual portfolio return** based on the recommended portfolio
Actual returns using the forecasted prices
Recommended portfolio
Used the program calculate the optimal protfolio based on actual returns
Recommended portfolio and actual returns with perfect information on prices
Comparison and Validation
Incrementally constrained the maximum investment per fund
Predicted and Actual Returns of 10 optional portfolios .

Figure 17. *The process outline for creating an optimal portfolios based on potential returns and actual returns.*

5.1 Initial Optimization Model

Before creating the optimization program using JavaScript, I first constructed my optimization model as shown in **Figure 18** below. This optimization model seeks to maximize the total return in dollars using the decision variable, x[i], or the number of shares of fund i to buy as well as the price and return (predicted or actual) of each fund. To create this model, I identified three constraints that should be included. The first was a risk constraint, which sought an average risk value or beta value in this case. Constraint 1 weights the difference between each fund's beta value and the desired portfolio beta value by the number of shares bought multiplied by the price of each fund. My calculation for this is shown below in **Figure 19**. The sum of these weighted differences must be less than or equal to zero, which means that the weighted average beta of the optimized portfolio must be less than or equal to the desired beta value. The remaining constraint ensure that the end portfolio does not cost more than the cash available and that the number of shares bought will be non-negative.



Let:

$n = number\ of\ funds$

$r[i] = return\ of\ fund\ i\ (i = 1..n)$

$p[i] = price\ (\$)\ of\ fund\ i\ (i = 1..n)$

C = total cash available to invest

$b[i] = beta\ value\ of\ fund\ i\ (i = 1..n)$

$B = desired\ beta\ value\ of\ portfolio$

$x[i] = number\ of\ shares\ of\ fund\ i\ to\ buy\ (i = 1..n)$

Maximize portfolio return:

$\sum_{i=1}^{n} p[i] * r[i] * x[i]$

Subject to:

constraint 1: $\sum_{i=1}^{n} x[i] * p[i] * (b[i] - B) \leq 0$

constraint 2: $\sum_{i=1}^{n} x[i] * p[i] \leq C$

constraint 3: $x[i] \geq 0,\ \forall i = 1..n$

Figure 18. *The initial optimization model used as a basis for the JavaScript program.*

$If\ n\ =\ 3\ and\ y[i]\ =\ x[i] * p[i]\ =\ dollar\ amount\ spent\ on\ fund\ i,$

$then,$

$$\frac{y1 * b1 + y2 * b2 + y3 * b3}{y1 + y2 + y3} \leq B$$

Multiply by $\boldsymbol{y1 + y2 + y3}$

$$y1 * b1 + y2 * b2 + y3 * b3 \leq B * y1 + B * y2 + B * y3$$

Subtract $\boldsymbol{B * y1 + B * y2 + B * y3}$

$$y1(b1 - B) + y2(b2 - B) + y3(b3 - B) \leq 0$$

Summation form

$$\sum_{i=1}^{3} yi(bi - B) \leq 0$$

$$\sum_{i=1}^{3} xi * pi * (bi - B) \leq 0$$

Figure 19. *The derivation of constraint 2 in the portfolio optimization model.*

5.2 Initial JavaScript Programming Application

Once I constructed the optimization model, I created a JavaScript project using IntelliJ IDEA Ultimate with a Gradle wrapper. Gradle allowed me to import the libraries necessary to create the optimization model within my program. The library I used, Apache Commons Math 3, may be found at commons.apache.org. This library allowed me to create a linear objective function object, a list of linear constraints as well as a non-negativity constraint, and an implementable Simplex Solver. This provided a simple translation from my written model to a usable program in JavaScript.

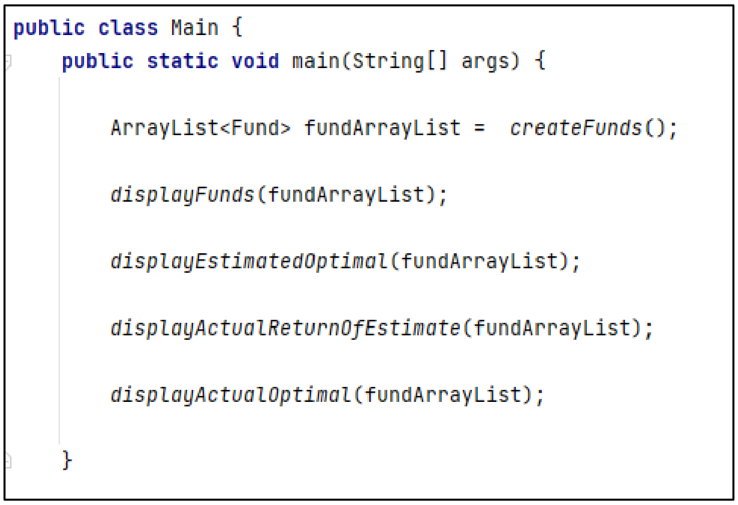
My program was created out of two classes, Main.java and Fund.java. The Main.java class was comprised of my main method and constructed methods, while the Fund.java class was an object class that specified Fund attributes, specified in **Table 14** below. The Fund.java class also included Fund object constructor method as well as getter and setter methods for the object attributes.

Table 10. *The attribute types included in the Fund object of the Java program.*

Attribute Name	Type
fundName	String
Ticker	String

riskValue	double
estimatedReturn	double
actualReturn	double
currentPrice	double

Within the main method of my Main.java class (**Figure 20**), I created an ArrayList for Fund objects using a constructed method called createFunds(). Then, I implemented four methods that display the fund information, optimal portfolio based on expected returns, actual returns of that portfolio, and the optimal portfolio based on actual returns. Each method called one or more other constructed methods.



```
public class Main {
    public static void main(String[] args) {

        ArrayList<Fund> fundArrayList =  createFunds();

        displayFunds(fundArrayList);

        displayEstimatedOptimal(fundArrayList);

        displayActualReturnOfEstimate(fundArrayList);

        displayActualOptimal(fundArrayList);

    }
```

Figure 20*. The main method within the Java progam including the sub-methods called.*

Figure 21 and **Figure 22** below feature the flow chart of methods called within the main method of my Main.java class. The methods in color are those that were called in the main method, and the methods in gray are the second layer constructed methods, or the methods that I called within the methods that are in color. Although most if not all of the second layer methods called a third layer of methods, none of these were constructed by me. All third layer methods were taken from the standard java library or the Apache Math Commons-3 library.

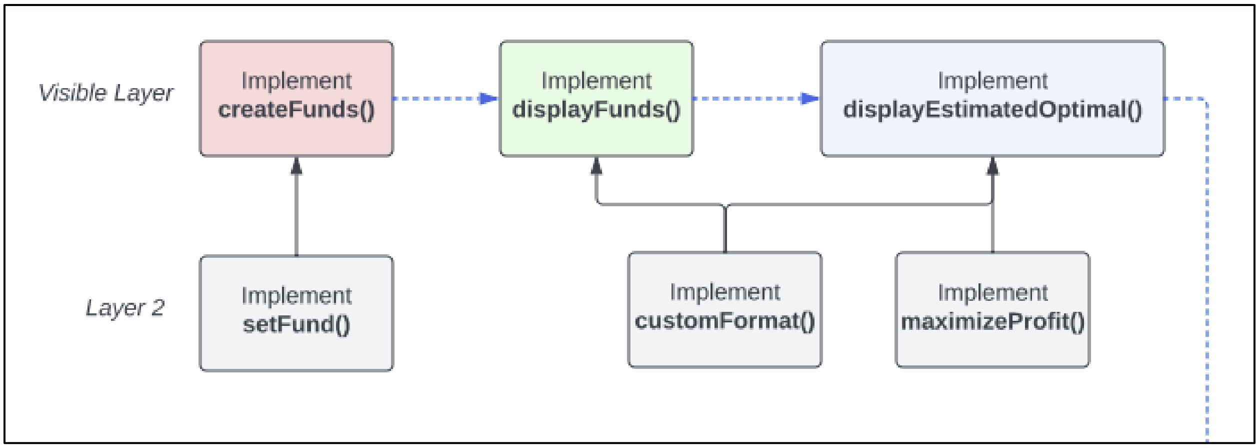


Figure 21. *The flow chart for three of the methods called within the main method of the Java program.*

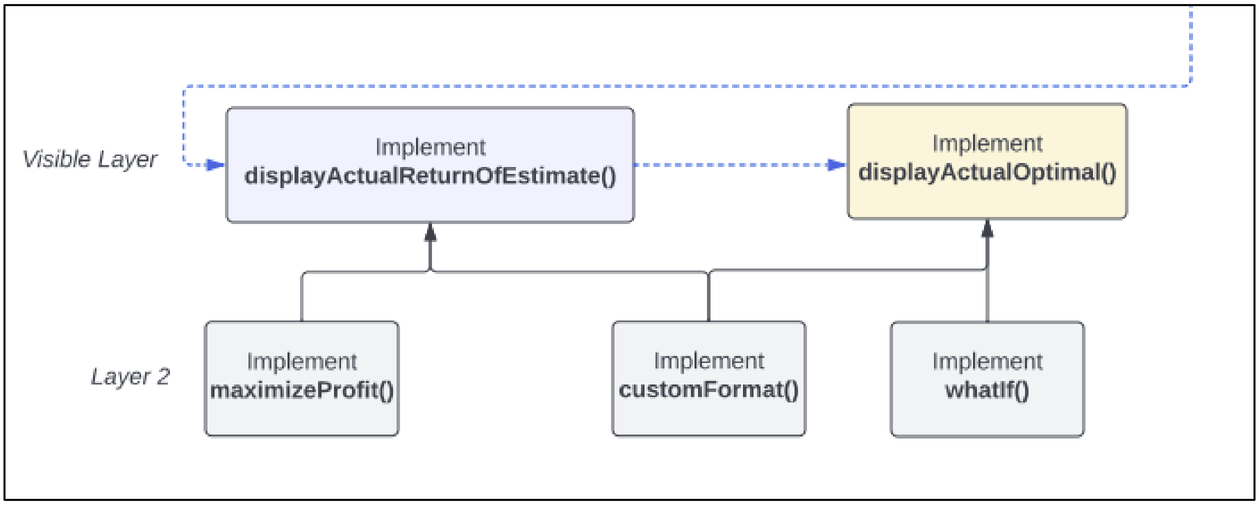

Visible Layer
Implement displayActualReturnOfEstimate()
Implement displayActualOptimal()
Layer 2
Implement maximizeProfit()
Implement customFormat()
Implement whatIf()

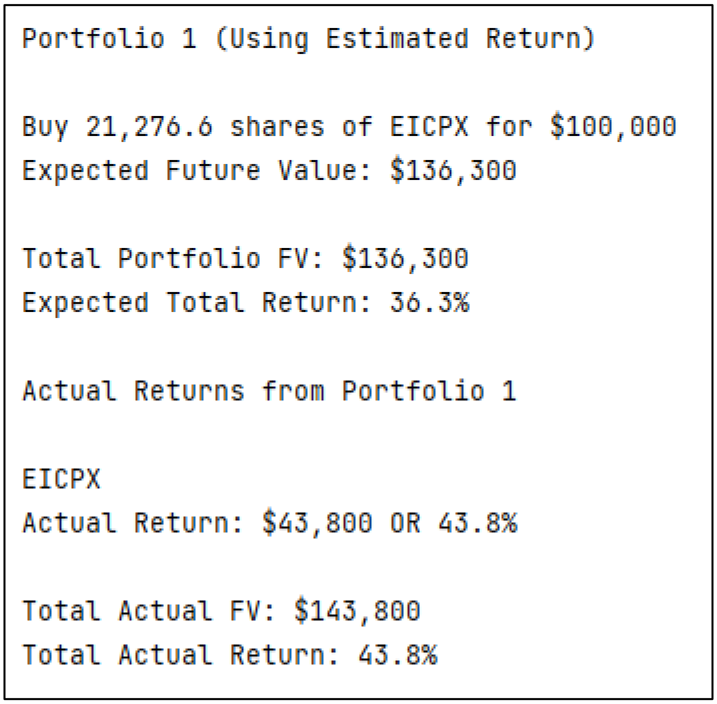
Figure 22. *The flowchart for two of the methods called in the main method of the Java program.*

Shown in Figure 21, the setFund() method was created so that I could set the attributes of each fund within a callable method. This in turn reduced the amount of code in my main method and made debugging simpler. The customFormat() method as shown is also used in the two methods in Figure 22 as well. I found this method using stackoverflow.com, which in turn uses a java library decimal formatter method and the format() method. While using the customFormat() method added an extra layer to the program instead of using the methods inside of it, it reduced the amount of code needed within other methods as I used the customFormat() for all number outputs written.

The maximizeProfit() method, which I used for both the displayEstimatedOptimal() method and the displayActualReturnOfEstimate() method, housed my optimization model. This method used the estimated returns, beta values, and prices of funds as well as the desired weighted average beta value and cash available. This method returned an ArrayList of Double values which stored the number of shares of each fund to purchase. Each index of the ArrayList corresponded to a specific fund in the ArrayList, fundArrayList. Like the maximizeProfit() method, the whatIf() method housed an optimization model and returned an ArrayList of Double values which stored the number of shares of each fund to purchase. However, the whatIf() method used the actual returns of each fund instead of predicted returns.

5.3 Java Program Results

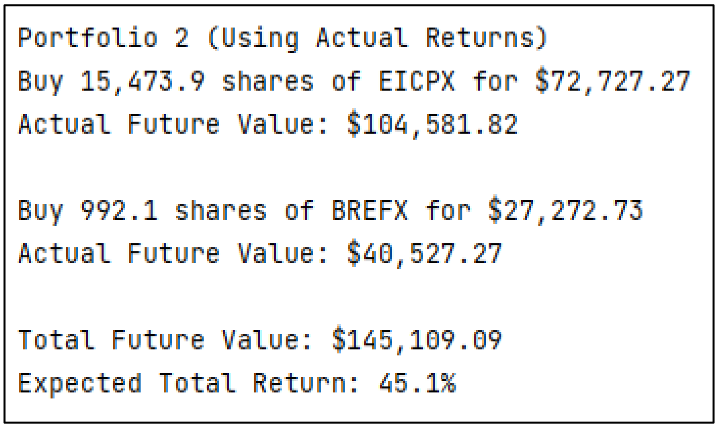
Using a desired average beta of 1.00 and the available cash amount of $100,000, I ran my model, which printed the results shown in **Figure 23** and **Figure 24**. The values of the unmentioned fund parameters are shown in the model verification in Figures 25, 26, and 27. The recommendation of portfolio 1 (optimized out of predicted returns) was to allocate all cash available to EICPX, which resulted in an expected return of $36,300 on top of the original investment. If an investor were to follow the instructions of this, they would have gained an actual return of $43,800 or 43.8%. The results of portfolio 1 may be attributed to EICPX's beta value and predicted returns. Out of the considered funds, EICPX ranked second in terms of desired beta at 0.94 and first in predicted return at 36.6%. Therefore, it makes sense that this was the recommendation because the fund with the best beta value of 0.81, VDIGX, only had a predicted return of 26.3%.



```
Portfolio 1 (Using Estimated Return)

Buy 21,276.6 shares of EICPX for $100,000
Expected Future Value: $136,300

Total Portfolio FV: $136,300
Expected Total Return: 36.3%

Actual Returns from Portfolio 1

EICPX
Actual Return: $43,800 OR 43.8%

Total Actual FV: $143,800
Total Actual Return: 43.8%
```

Figure 23. *The Java program output fo rthe optimal portfolio based on potential fund returns.*

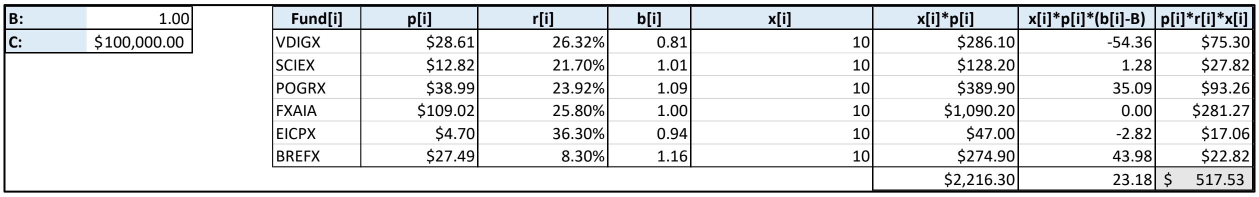
While my programs recommendation for portfolio 1 only included EICPX, the recommendation of portfolio 2 included EICPX and BREFX. Based on actual fund returns, portfolio 2 allocated 72.7% of the cash available to EICPX and the rest to BREFX. Like portfolio 1, these results appear to make sense because EICPX still has the second highest actual return at 43.8% and its relatively low beta value of 0.81. BREFX had the highest actual return of 48.6%, but the program did not allocate the majority of available cash to it because of its relatively high beta value of 1.16, which was the highest out of the funds.



```
Portfolio 2 (Using Actual Returns)
Buy 15,473.9 shares of EICPX for $72,727.27
Actual Future Value: $104,581.82

Buy 992.1 shares of BREFX for $27,272.73
Actual Future Value: $40,527.27

Total Future Value: $145,109.09
Expected Total Return: 45.1%
```

Figure 24. *The Java program output for the optimal portfolio based on actual fund returns.*

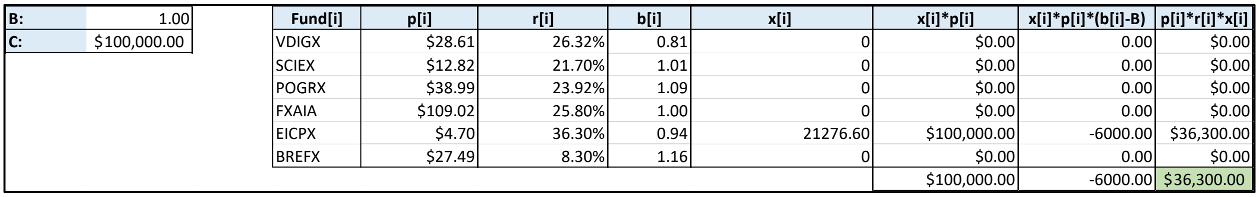
After completing and running my model, I chose to verify the results using Microsoft Excel. Specifically, I used the Solver tool with Simplex LP to maximize the objective function. In **Figure 25** is shown my initial fund parameter table and calculations in excel. Within this initial model, the shares to buy of each fund were preset to 10, and the returns were set to the predicted returns of each fund. Within Solver I selected to have it maximize the sum of the total return from dollars allocated to each fund (p[i]*r[i]*x[i]), which is shown in gray. To constrain the model, I set the sum of dollars spent on each fund (x[i]*p[i]) to be less than or equal to the cash available, and I set the sum of each funds value for x[i]*p[i]*(b[i]-B) to be less than or equal to 0. Also, I added non-negativity constraints to each x[i]. The results of implementing solver with these constraints using predicted returns is shown in **Figure 26**, and the results using actual returns is shown in **Figure 27**.



B:	1.00
C:	$100,000.00

Fund[i]	p[i]	r[i]	b[i]	x[i]	x[i]*p[i]	x[i]*p[i]*(b[i]-B)	p[i]*r[i]*x[i]
VDIGX	$28.61	26.32%	0.81	10	$286.10	-54.36	$75.30
SCIEX	$12.82	21.70%	1.01	10	$128.20	1.28	$27.82
POGRX	$38.99	23.92%	1.09	10	$389.90	35.09	$93.26
FXAIA	$109.02	25.80%	1.00	10	$1,090.20	0.00	$281.27
EICPX	$4.70	36.30%	0.94	10	$47.00	-2.82	$17.06
BREFX	$27.49	8.30%	1.16	10	$274.90	43.98	$22.82
					$2,216.30	23.18	$ 517.53

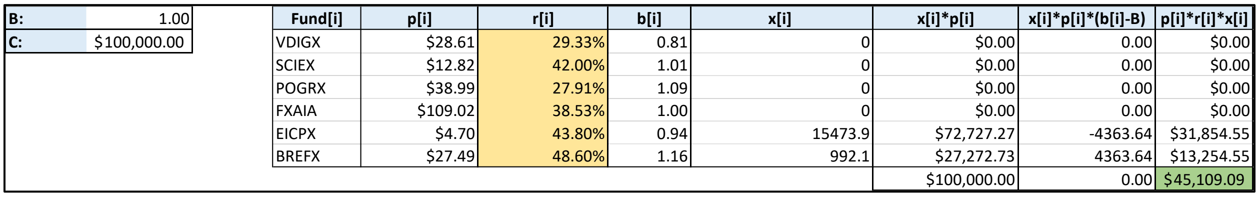
Figure 25. *The Excel optimization setup before implementing the Solver tool.*



B:	1.00
C:	$100,000.00

Fund[i]	p[i]	r[i]	b[i]	x[i]	x[i]*p[i]	x[i]*p[i]*(b[i]-B)	p[i]*r[i]*x[i]
VDIGX	$28.61	26.32%	0.81	0	$0.00	0.00	$0.00
SCIEX	$12.82	21.70%	1.01	0	$0.00	0.00	$0.00
POGRX	$38.99	23.92%	1.09	0	$0.00	0.00	$0.00
FXAIA	$109.02	25.80%	1.00	0	$0.00	0.00	$0.00
EICPX	$4.70	36.30%	0.94	21276.60	$100,000.00	-6000.00	$36,300.00
BREFX	$27.49	8.30%	1.16	0	$0.00	0.00	$0.00
					$100,000.00	-6000.00	$36,300.00

Figure 26. *The Solver tool output for the optimal portfolio based on potential fund returns.*



B:	1.00
C:	$100,000.00

Fund[i]	p[i]	r[i]	b[i]	x[i]	x[i]*p[i]	x[i]*p[i]*(b[i]-B)	p[i]*r[i]*x[i]
VDIGX	$28.61	29.33%	0.81	0	$0.00	0.00	$0.00
SCIEX	$12.82	42.00%	1.01	0	$0.00	0.00	$0.00
POGRX	$38.99	27.91%	1.09	0	$0.00	0.00	$0.00
FXAIA	$109.02	38.53%	1.00	0	$0.00	0.00	$0.00
EICPX	$4.70	43.80%	0.94	15473.9	$72,727.27	-4363.64	$31,854.55
BREFX	$27.49	48.60%	1.16	992.1	$27,272.73	4363.64	$13,254.55
					$100,000.00	0.00	$45,109.09

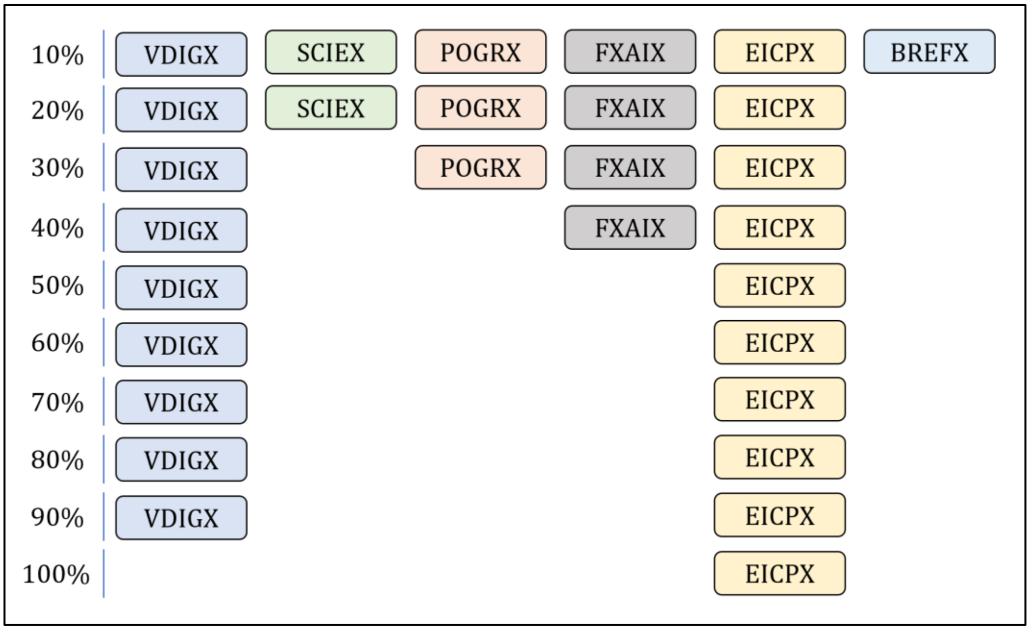
Figure 27. *The Solver tool output for the optimal portfolio based on actual fund retruns.*

Using Excel's Solver tool proved useful for verifying that my Java program operated correctly as the optimal allocation of cash available using Solver matched that of my program for both portfolio 1 and 2. While it may seem more practical to use Solver for my purposes instead of constructing a Java program, using Excel in this way lacks the potential for scaling. Furthermore, using a Java program allows me to give user abilities for those who may not know how to use solver possibly through a user-interface. This in turn would allow investors to place excel files of fund information and automatically calculate an optimal portfolio.

5.4 Reconsidering the Needs of an Optimized Portfolio

Although the optimization program I created solved for the optimal solution in terms of the required constraints and the desired maximization of return, it did not fully capture the needs of a portfolio. Specifically, an optimized portfolio made up of one or two funds should not be considered as diversified, and the purpose of this study was to create a portfolio that spreads risk out across investments while producing a substantial return. Furthermore, if I were to include hundreds of funds instead of just six funds, it would seem arbitrary to have a result of just investing in one fund. Therefore, I chose to attempt to constrain the maximum investments possible for each fund.

To add these needed constraints, I reviewed the available functionality of the Apache Commons Math-3 library. However, this library requires constraints to be added in the form of an array, where each index contains the coefficient of each of the decision variables being added together. Therefore, I was unable to incorporate a constraint for each of the decision variables, so to address this, I used the same Excel solver processes as used to verify my original optimization models. Using Excel's solver tool, I created a table like the one shown in Figure 25. For ten iterations, I constrained the total dollar value invested in each fund to be less than a certain value. I chose the values as a percentage of the available capital, so the values ranged from 10% to 100% and were incremented by 10%. The resulting funds selected at each increment are shown below in **Figure A**, and the comparison of each increment in terms of number of funds selected, expected performance, and actual performance are shown in **Figure B.**



10%	VDIGX	SCIEX	POGRX	FXAIX	EICPX	BREFX
20%	VDIGX	SCIEX	POGRX	FXAIX	EICPX	
30%	VDIGX		POGRX	FXAIX	EICPX	
40%	VDIGX			FXAIX	EICPX	
50%	VDIGX				EICPX	
60%	VDIGX				EICPX	
70%	VDIGX				EICPX	
80%	VDIGX				EICPX	
90%	VDIGX				EICPX	
100%					EICPX	

Figure 28. *The resulting fund selections for each maximum allocation constraint level.*

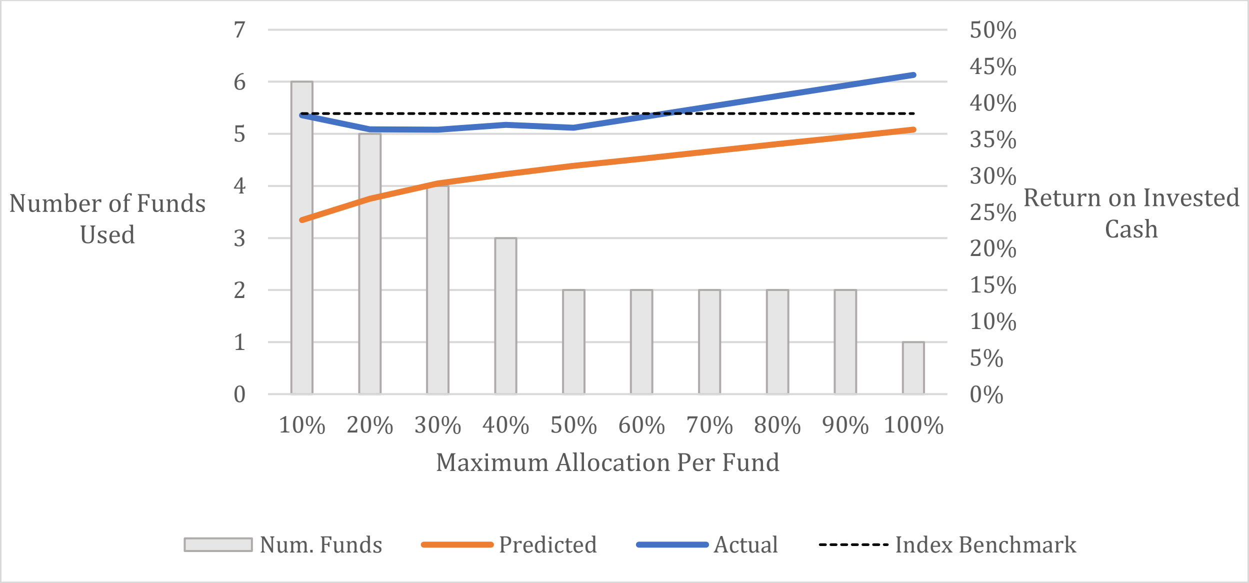


Figure 29. *The resulting performances of the portfolios created at each maximum allocation constraint.*

As shown in Figure B, the return on invested cash does not significantly increase until the maximum allocation per fund exceeds 60% of the available cash. Among the portfolio options under this threshold, the portfolio with six funds and the portfolio with two funds at 60% maximum allocation appear to have the greatest return on invested cash. For the sake of diversification, one's initial inclination might be to invest in the portfolio with six funds, they must consider something. This portfolio could only invest up to 60% of the total available cash amount. While it had a higher return percentage than the following four portfolios, this was based on the cash invested. Therefore, its actual return in dollars was about $23,700, whereas the return of the following portfolio was $36,300 a lower return percentage. However, if an investor wanted a high degree of diversification and maintain a decent total cash return, it would be advisable to equally invest in each fund and use the total amount of available cash. Except for BREFX, all funds received an equal investment for the first portfolio, and scaling this to a total of $100,000 invested could provide a high cash return. BREFX was not assigned the full amount by the model because of its significantly higher beta value. However, the difference was nominal, and in a practical environment, it would not be necessary to reduce its investment.

Another insight that may be drawn from Figure B is that the predicted return consistently increases as the constraint maximum increases, whereas the actual performance decreases and then increases. When decisions are made based off of the predicted returns, it would have been difficult to make the recommendation of using an equally allocated portfolio because there would have been no way to know the actual outcome. Therefore, the investment decision would require a more in-depth trade-off analysis between diversification and return.

6. Discussion

6.1 Summary of Results and Insights

Using linear regression forecasting and portfolio optimization, I created two portfolio that maximized the potential and actual returns of fund selections. As per one of my stated goals, this portfolio outperformed the index fund used in selection and as a comparison, VDIGX. Had an individual investor utilized my initial recommended portfolio consisting of a 100% investment in EICPX, they would have gained a 43.8% return. However, the predicted return of this recommendation was in fact 36.3%. While this may have been highly profitable, it was based on relatively inaccurate forecasts as the actual return was approximately 21% higher than the predicted return. Although the actual was higher than the predicted, this may not always be the case in terms of variation and creating a better forecasting model would have reduced the likelihood that actual returns were significantly lower as well.

6.1.1 Selection of Funds and Macroeconomic Factors

During the selection process of choosing funds and factors, it became apparent that the forecasts I made were only as good as the factors chosen as well as the time periods in which factors were used. After selecting the macroeconomic factors, I noticed that the interest rate factor was creating inaccurate results as I attempted to forecast prices with factors in the same period. Therefore, I chose to not include this factor in future forecasts. Although I would have been able to correct his factor as I did with the forecasts used in my portfolio optimization phase, it was apparent that the model was assigning high importance to the interest rate factor despite the inaccuracy it caused. Because of this, I realized that it is also important to fully understand the factors selected and their range of values. The other piece of evidence that led me to the conclusion of factor selection importance was my use of step wise regression. Because the model clearly selected factors of which it deemed important to the price response, it is important to select the best input factors as well. While the model will not assign factor importance perfectly, those creating it should clearly pick relevant factors that they understand.

The importance of selecting the correct funds for consideration was evident both in my forecasting stage as well as my portfolio optimization stage. When considering a fund's use in forecasting, investors need to consider how statistically predictable the fund has been in the past in relation to macroeconomic factors. Although I considered how funds might respond to factors, I had not considered how well individual funds were predictable in the past. While someone may attempt to understand this in a complex way, I believe, at least for the individual investor, it is more practical to observe the graphical volatility of a fund's price as well as find metrics that quantify the risk and volatility. While this may not necessarily provide an indication of the predictability of a fund, it does provide a practical avenue to selecting more stable funds.

Regarding the optimization phase of my study, I realized that the selection of funds as well as the number of funds is important to diversifying a portfolio. Specifically, during my initial optimization, my model selected only one fund to invest in, EICPX based on its expected return and beta value. While this may have been a good recommendation within a vacuum where only my selected factors matter, it does not capture the complex risk associated with investing all capital into one fund. For example, while all fund categories might be affected by a certain global event such as the pandemic or the invasion of Ukraine, they will undoubtedly be affected differently. Because of these events, it is evident that individual investors must practice wisdom in selection of their funds in addition to the statistical techniques used. Furthermore, for an optimization model to correctly choose the optimal mixture of investments, it must have the best possible options. Without careful consideration of inputs, the output of such a process will not fulfill the potential of an investor's available capital.

Not only did I learn that the selection of funds and factors are important, I realized that the selection process as well as the justification of such is important. Specifically, without a properly defined process, the funds a person selects might be arbitrary. Although an arbitrary selection may result in a profitable investment, an investor's ability to predict this probability is most likely also arbitrary. Furthermore, to bring the selection process of funds to a larger scale, it must be clearly defined to achieve consistency.

6.1.2 Forecasting Fund Prices

As I created the price forecasts, I learned about limitations of my models. One such limiting factor was the type of model, which was linear. While a linear model provides quick processing times and is computationally inexpensive to produce, it fails to capture the complex relationships between macroeconomic factors and fund or stock prices. Assuming a linear relationship between factors greatly limited the complexity that I needed but also reduced the characterization of factor relationships that I could achieve. Furthermore, the relationships between macroeconomic factors and fund prices may change with respect to time. While this may have been a problem if I applied the linear model to forecast more than one year ahead, it may have not been a significant limiting factor to my models.

To address the forecasting error present in the projected returns, multiple other forecasting methods could be used. A random forest learning model would be a suitable replacement for a linear regression model because it captures complex relationships better using a decision-tree based modelling. Furthermore, it does not come at a significantly larger computational expense compared to linear regression. While other decision tree-based models, such as XG-Boost models, require parameter tuning to gain higher accuracy, the incremental benefit of using a random forest model over a linear regression model most often outweighs the added time and effort required.

Another limitation of my models were the time frames of training data used as well as unquantified global events. Except for two funds, the eight-year training period model was the selected model across most funds. Although the longer time frame provided more data to capture the relationships between factors and prices, it may not have been long enough. As discussed, the dramatic increase in unemployment rate in my testing set may have caused inaccurate forecasts because my training set did not include the extreme values shown in the testing set. While the solution may be to simply extend the training set length to include a decade or more of data, this may not capture current market relationships. The American economy has changed dramatically in the last three years due to the pandemic, so it may be difficult to assume that the responses of prices to factors today would be the same as three years ago.

While I do not completely understand the relationships between prices and macroeconomic factors, it is safe to say that these relationships likely change over time. Therefore, one future opportunity may be keeping weighting the model training data based on age. Conceptually, applying quantitative importance based on the age of data to forecasting models may yield models that are more grounded on recent relationships. This in turn would continue to consider older relationships, but it would recognize the change in these relationships better to create a more accurate model of current ones. Another future area of opportunity would be quantifying global events. While it is difficult to say how this may take shape, it would be useful for investors to determine how different types of events affect fund prices.

The pandemic represented an unquantified event that created prices responses that could not be accurately forecasted at least by my models. These events, which may include the 2008 collapse of the housing market, make it difficult to assume steady price relationships over a long period of time. The data I used to test the models came from the dramatic economic growth following the initial pandemic, which may explain the high fund returns. This phenomenon calls into question of whether we should forecast these funds on a 10-year basis or 1 year basis. If a 10-year basis is used, we would be able to average returns to reduce the influence of extreme price changes. However, it may be useful to reevaluate on a shorter term to reduce the risk of keeping a failing investment.

While someone may create an accurate forecast based on steadily changing macroeconomic factors, a forecast's accuracy is highly susceptible to the complex global economy. Within the past six months, the world has seen the trailing ends of a pandemic, global supply chain issues, and a foreign invasion. I learned that no forecast can accurately predict global events such as these, and the effects they have on macroeconomic factors such as inflation or CPI. This unpredictable change in the factors used in my forecasting also provided insight into the timing of factors used. Specifically, my first forecasts were done with the factor values from the same time as the forecasted prices. Although this forecast proved accurate after I removed the interest rate factor, it assumed I would be able to accurately predict the macroeconomic factors as well. I learned that this assumption was invalid in my models because the factors did not have a steady variation, and therefore, I could not easily foresee their future changes. Because of this, it appeared far more practical to predict future prices with current factors.

6.1.3 Creating Optimized Portfolios

As a result of implementing an optimization program in Java, I gained a better understanding of the limitations of a simpler model and its recommendation. As discussed, allowing the model to produce the optimal portfolio resulted in it choosing only one fund based on potential returns. However, one characteristic that I missed was diversification. While simpler models produce simpler results, they cannot fully prepare an investor for the future because they may not adequately spread risk. While I had the luxury of seeing the actual return of my portfolio recommendation, investors will not, and they therefore must spread risk across multiple investments. Furthermore, I learned that it is necessary to place constraints on an optimization model to gain a more practical optimal portfolio.

Once I had placed the constraints on the maximum investment in each fund, it appeared that the potential and actual return did not increase substantially as the portfolio complexity (number of funds) decreased. In fact, the actual return percentage took a dip before increasing slowly. Based on my results, I would recommend creating a new portfolio based on the portfolio with six funds. This portfolio had one of the highest return percentages among the more diversified portfolios, but it had the lowest cash return because the total maximum investment was constrained by the model. Practically, this portfolio equally allocated the investment amount per fund. Therefore, investing all of the available ash, equally allocated per fund would yield both a high percentage and high cash return while maintaining a balanced diversification. However, for individual investors, they must make this decision given their willingness to give up potential returns for incremental diversification. Furthermore, they would not have had the information on actual returns, and they therefore would have had to make a decision given the predicted return curve. This conclusion undoubtedly sums up my study because no recommendation should be given to an individual investor as the optimal one for them. Instead, individual investors should receive a series of recommendations that vary along risk and return, and they must be able to make the decision of given the information provided to them.

6.2 Application to the Individual Investor

Regarding individual investors, I have learned through this study that they should identify a broad spectrum of possible funds to invest in, select simple factors to make price predictions, be hesitant to respond eagerly to price forecasts, and understand how much return they are willing to give up for increased portfolio diversification. When investors consider the possible inputs of their portfolios, they should take careful consideration of the number involved. Although investors may not forecast fund prices themselves, it is worth noting that they should not limit themselves to a fixed number. While I was able to create a more complex portfolio that balanced risk and return, the maximum number of funds that I could spread risk across was six. However, if I had included more fund categories and more funds within each category, I could have tuned my final portfolio to possibly gain a higher return at the same level of diversification. Given the varying beta values of unconsidered funds, an optimization model might choose the portfolio selection differently. Therefore, it would be in the investor's favor to have more than enough options than too few.

When selecting factors, it was clear that macroeconomic factors as well as fund prices are heavily entangled into the global economy. Unless an individual is a data scientist, it would be difficult for them to interpret the change of several factors at once as well as the affect they have on a fund. Therefore, for the purposes of an individual investor, I believe it would make more sense to focus on a select few factors that are not highly volatile. Although an individual investor might want to sort through the complex relationships of several factors, far beyond the number I used for this study, they should begin with a simpler set of factors that they understand. If an investor does not understand their selected input factors, they will most likely not understand the output forecast for a fund. When responding to a price forecast, individual investors should be wary of the immediate desire to invest. Although the actual returns of all my selected funds were higher than their constituent predicted returns, this prediction error may not always point in the positive direction, so investors must be cautious on how they respond to forecasts. No investment decision regarding funds will ever be made with perfect information, but investors may combat this problem by finding credible sources and understanding the forecasts from them.

Individual investors must also understand their price of risk or the level of risk their willing to take on for a certain return. It was clear that my unconstrained portfolio lacked diversification, and if an investor chose the recommendation of such a portfolio, they would incur a heavy risk. However, after constraining my model, my output provided a series of portfolios varying by diversification and return. While I may have been able to make a sound recommendation, it is up to individuals to understand how much return they are willing to give up to distribute risk across multiple investments. At the core of being an investor, individuals must be aware of their own tolerance for risk. While mutual funds are not the most risk-prone investment product on the market, individuals will ultimately hold responsibility for investment decisions. Therefore, individual investors must understand their own financial boundaries as well as constraints before attempting to understand a high-dollar investment or acting upon it.

As it applies to my own investment knowledge, I now understand the importance of identifying where predictions come from. Although I could forecast every fund that I consider investing in, it might not be a wise use of my time. However, using credible source such as Morningstar may give me an edge in my investments. Based on the optimization results of my research, I would like to identify my own tolerance to risk and be more intentional to find the right balance between risk and return. Finally, this study has provided me the confidence to invest strategically, and I believe that is what stops many individual investors. While relying on a financial advisor is by no means an undesirable situation, with the right knowledge, individual investors such as myself can invest strategically without the fees that come with an advisor. Furthermore, I now know that I have responsibility for the decisions made on my behalf, so it makes sense to understand and take part in those decisions.

Appendix

A. Summary of the Selection of Forecasting Models for POGRX, FXAIX, BREFX, and SCIEX

Shown in **Table 8** below are the linear regression equation coefficients obtained from each of the 8 training periods for the PRIMECAP Odyssey Growth Fund (POGRX). The R-square values are shown as well. For POGRX, the R-square value of the eight-year training period was the highest at 85.25%, which was almost 7 percentage points above the next highest. Therefore, I chose to only plot the forecast of the eight-period training set, which is shown in **Figure 30** below. My reasoning for this was based on the results of forecasting VDIGX, in which the eight-year training period forecast produced the smallest error.

Table 11. *The regression equation constants and R2 values for the 8 step wise regression models created for POGRX.*

Training Period (Years)	Equation Constant	Current Price Coefficient	Unemployment Rate Coefficient	CPI Coefficient	Debt Coefficient	Performance Coefficient	Exports Coefficient	R^2
1	605.7	0	0	0.162	-5.02	-1.303	03.918	69%
2	100.2	-0.434	0	-0.171	0	-0.841	0	39%
3	322.6	-0.292	-14.52	0.855	0	-0.551	0	30%
4	182.2	0.335	-7.40	0	0	0	-1.009	48%
5	82.86	0	-6.673	0	0	0	-0.1416	66%
6	-60.6	0	0	0.496	0.785	0	-0.3320	71%
7	-64.0	0	0	0.487	0.828	-0.1802	-0.2930	78 %
8	-53.7	0	0	0.372	1.490	-0.2075	-0.2559	85%

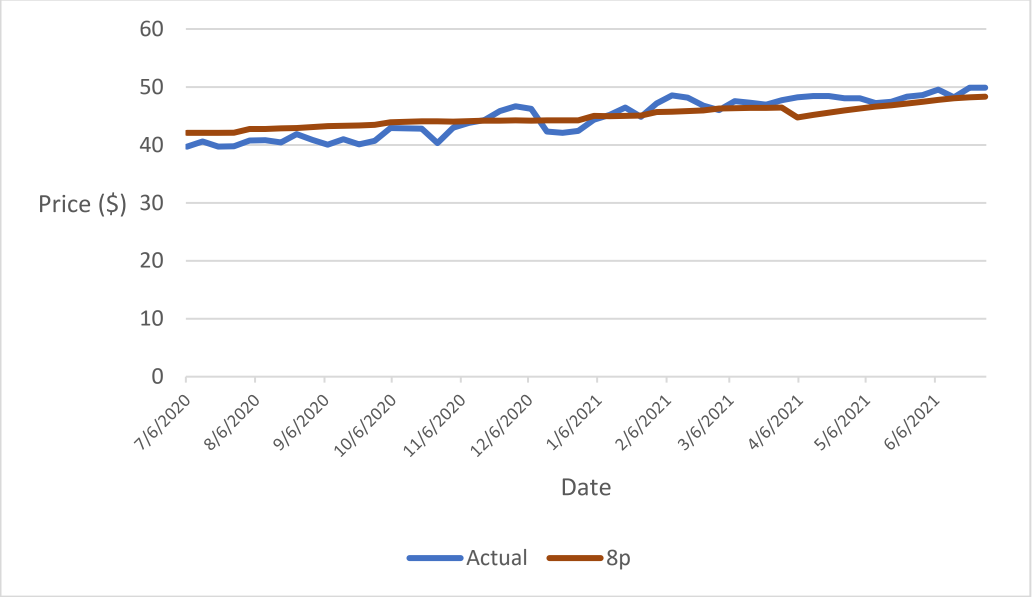


Figure 30. *The forecasted price for the 8-year training period model for POGRX.*

From Figure 30, the 8p forecast appears to be accurate relative to the actual prices during the prediction period, and with a 3.87% MAPE, the eight-period forecast has about the same forecasting error as the chosen forecast for VDIGX. Therefore, I chose to use the eight-year training period of POGRX for my portfolio optimization phase.

Shown below in **Table 9** are the linear regression equation coefficients obtained from each of the 8 training periods for the Fidelity 500 Index Fund (FXAIX) as well as the R-square values for each training period. Except for the one-year training period, the R-square values increased as the number of training periods increased, up to 93.97% for the eight-year training period. Like my approach with VDIGX, for FXAIX I used the training periods with the highest R-square values to test their forecast against actual price data during my testing period. For FXAIA, I used the top three, which were the eight, seven, and six-year training periods, and their forecasts are shown in **Figure 31** below.

Table 12. *The regression equation constants and R2 values for the 8 step wise regression models created for FXAIX.*

Training Period (Years)	Equation Constant	Current Price Coefficient	Unemployment Rate Coefficient	CPI Coefficient	Debt Coefficient	Performance Coefficient	Exports Coefficient	R^2
1	1061	0	0	0	0	-2.329	-7.50	54%
2	15.5	-0.650	0	0	6.95	0	0	33%
3	-21.6	0	0	0	5.851	-0.824	0	48%
4	101.0	0	-8.88	0	5.223	0	-0.601	76%
5	-101.7	-0.385	-7.06	0.642	5.01	0	0	86%
6	-220.9	-0.3599	0	0.832	6.943	0	0	88%
7	-189.4	-0.2636	0	0.653	7.151	0	0	91%
8	-180.7	-0.2135	0	0.591	7.290	-0.281	0	94%

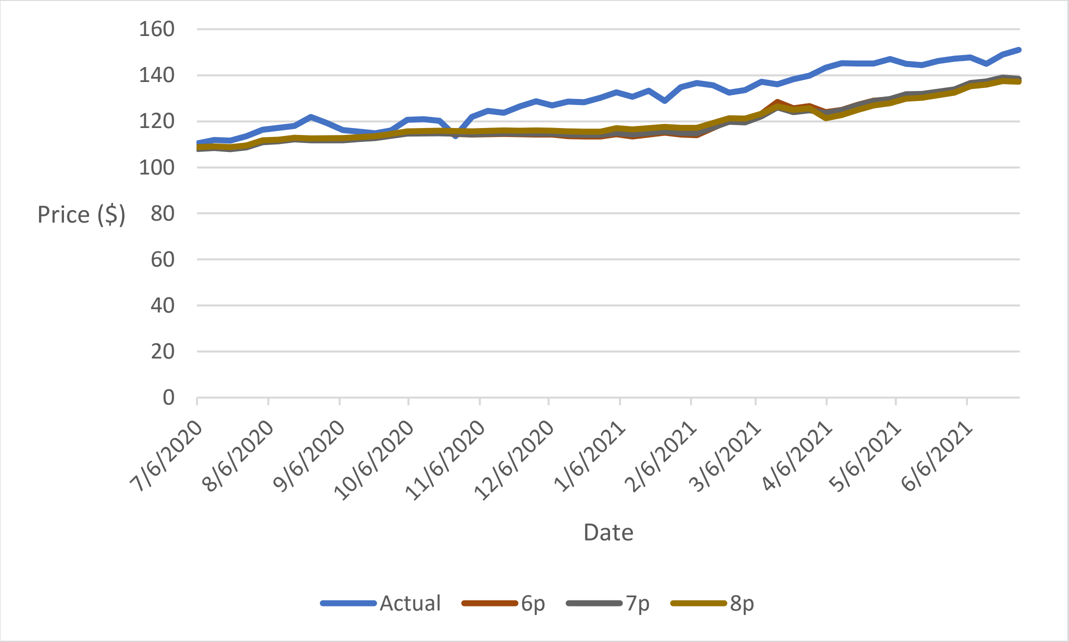


Figure 31. *The forecasted prices for the 8, 7, and 6-year training period models for FXAIX.*

As shown in Figure 31, the 6p, 7p, and 8p forecasts observably performed similarly. In fact, all share the same active regression factors with the exception of the eight-year training set equation including the performance factor. The graph alone was not sufficient to determine which forecast would be the correct one to continue with. Practically, any of the funds would have been sufficient to use for my portfolio optimization, but I chose the forecast using an eight-year training set because it had an MAPE of 7.88%. This was lower than the MAPE's of the 6p and 7p forecasts, which were 8.20% and 8.24% respectively.

Shown below in **Table 10** are the linear regression equation coefficients obtained from each of the 8 training periods for the Baron Real Estate Fund (BREFX) as well as the R-square values for each training period. Like EICPX, the R-square values for BREFX do not point to any sign of an increasing or decreasing trend as the training period increases. However, the one, three, and eight-year training period regression models had noticeably higher R-square values. Therefore, I chose to plot their forecasts against the actual prices of BREFX, as shown in **Figure 32**.

Table 10. *The regression equation constants and R2 values for the 8 step wise regression models created for BREFX.*

Training Period (Years)	Equation Constant	Current Price Coefficient	Unemployment Rate Coefficient	CPI Coefficient	Debt Coefficient	Performance Coefficient	Exports Coefficient	R^2
1	426.2	0	-9.49	0	-3.55	-0.793	-2.255	48%
2	66.5	0	0	0	0	0	-0.325	6%
3	91.00	-0.2207	0	0	0	-0.350	-0.4673	46%
4	42.68	0	0	0.2829	0	0	-0.698	18%
5	-94.2	0	5.068	0.6629	0	-0.2369	-0.5229	26%
6	23.42	-0.2281	0	0	0.4062	0	0	9%
7	37.23	0	-0.656	0	0	0	-0.0679	22%
8	24.15	0.2198	-0.855	0	0	0	0	51%

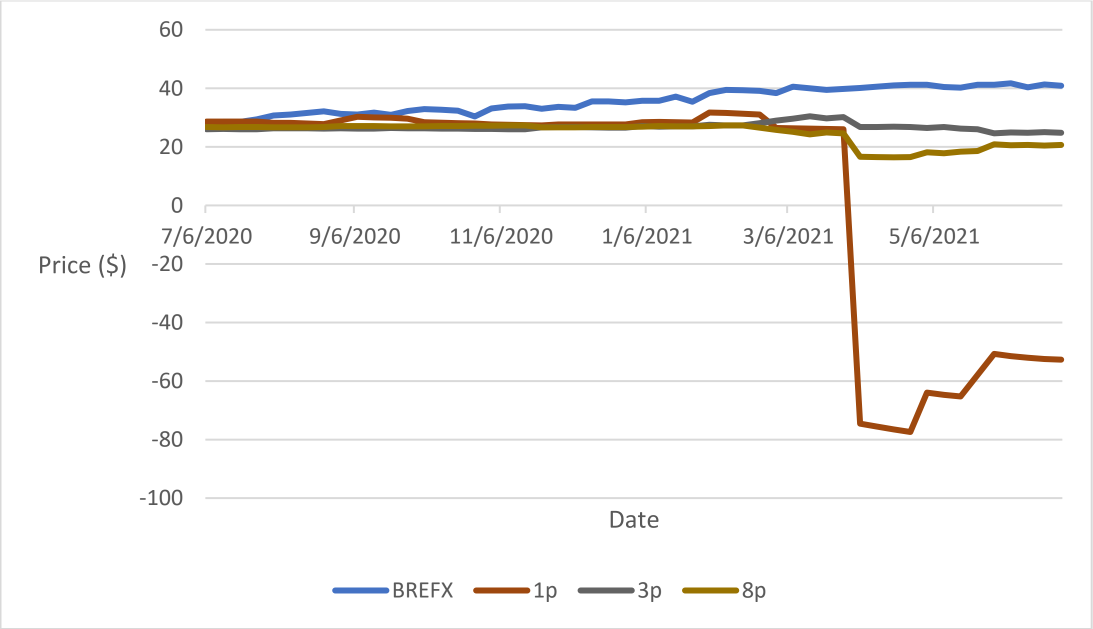


Figure 32. *The forecasted prices for the 8, 3, and 1-year training periods models of BREFX.*

As shown in Figure 32, all forecasts tended to decrease with time with a dramatic decrease of 1p's forecast around April of 2021. Like EICPX, I wanted to determine if the dramatic decrease in the forecasted price for 1p was due to dramatic macroeconomic factor changes. As shown in Table 10, 1p is dependent on the unemployment rate, which I determined had a dramatic shift, and it has a large negative coefficient for this factor. Furthermore, I determined that the performance index experiences a dramatic change around the same time that the unemployment rate does, and 1p is dependent on the performance index. The time series plot of the unemployment rate and performance index are shown in **Figure 33**.

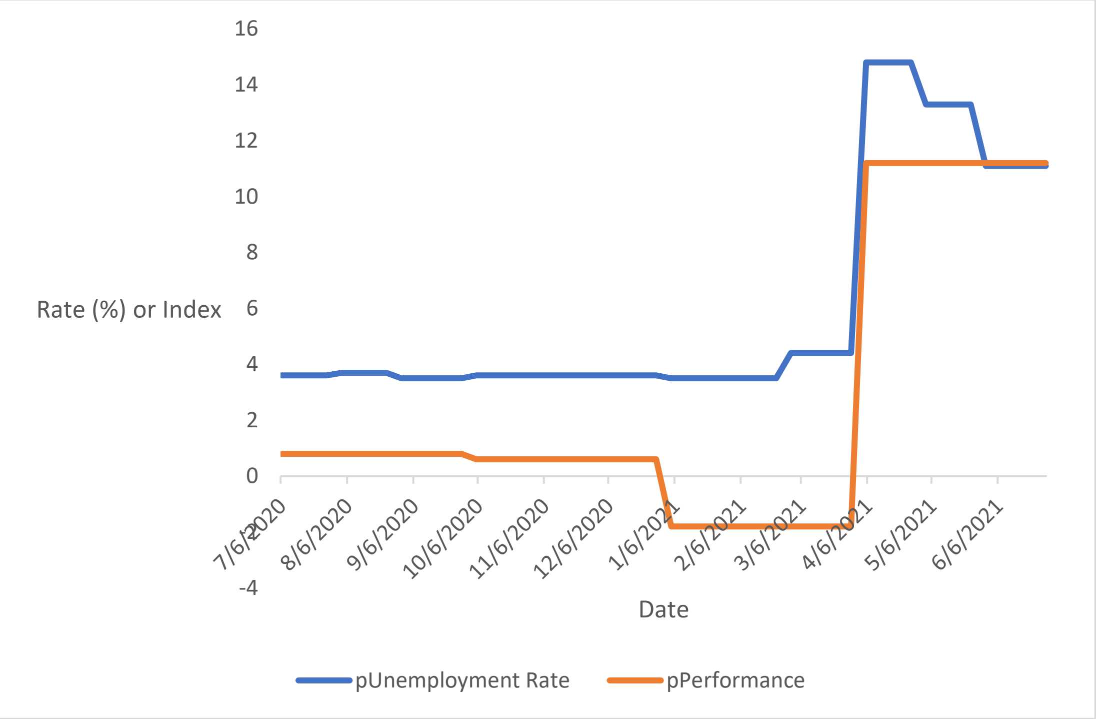


Figure 33. *The unemployment rate and performance index values used in the testing period.*

To choose a forecast, I used a similar approach for altering my models as I did for EICPX. By using the Microsoft Excel forecasted unemployment rate values shown in Figure 15 and new forecasted performance index values (**Figure 34**), I calculated new forecasts for the 1p, 3p, and 8p models. The plot of these new forecasts is shown in **Figure 35** below.

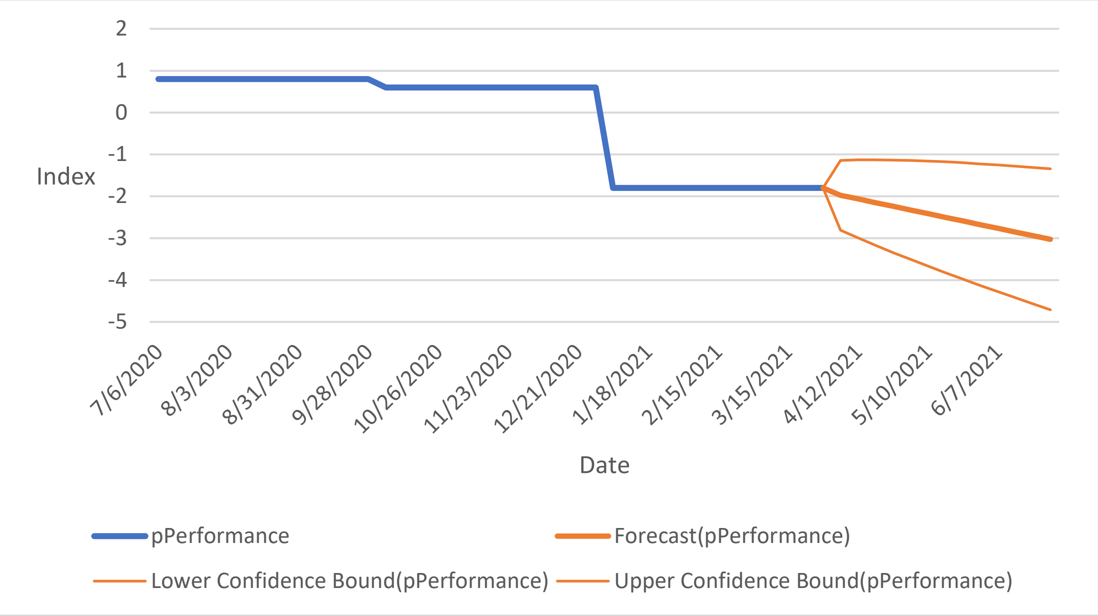


Figure 34. *The plot of the stabilized performance index values using Excel's forecasting tool.*

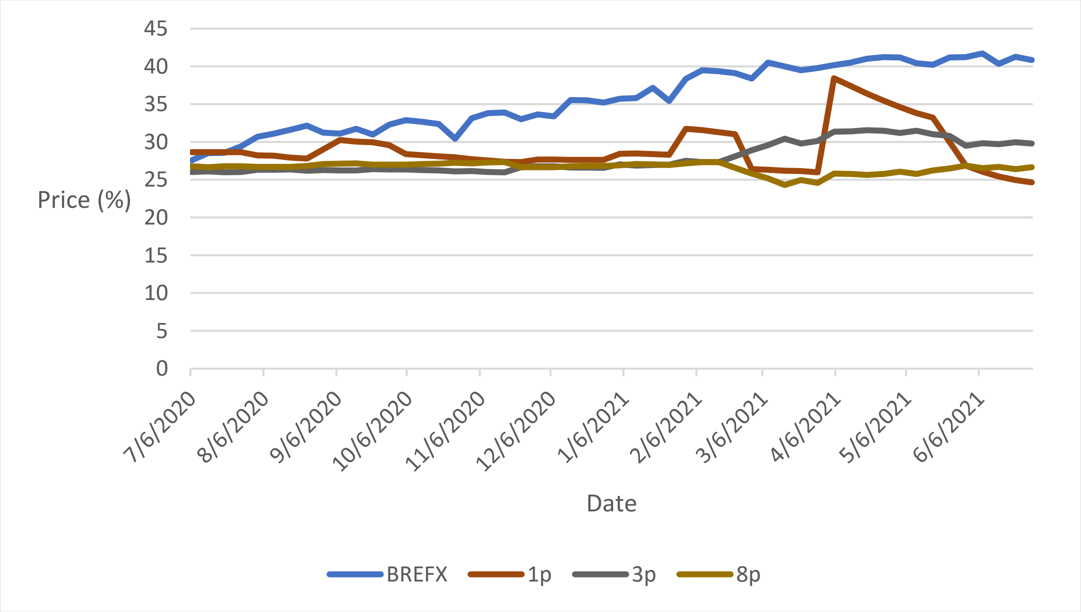


Figure 35*. The forecasted prices for the 8, 3, and 1-year training period models of BREFX with altered unemployment rate and performance index values.*

As shown in Figure 35, the forecast that I had targeted in my altered model, the 1p forecast, did not experience a dramatic decline in its altered forecast. However, after January of 2021, the forecast observably experiences high variation in predicted value. While the 1p forecast resulted in a 17.9% MAPE, the 3p forecast resulted in a 7.2% MAPE. Although my intention for altering the model by forecasting macroeconomic factor values was to improve the 1p forecast, the 3p forecast improved. Using an altered model, the 3p forecast predicted increasing returns instead of decreasing returns, and the MAPE decreased from 24.5% to 7.2%. Therefore, I chose the altered 3p forecast model to use in my portfolio optimization.

Shown in **Table 11** are the linear regression equation coefficients obtained from each of the 8 training periods for the Hartford Schroders International Stock I Fund (SCIEX) as well as their respective R-square values. Although the R-square values do not vary as dramatically as previous funds, they appear to decrease as the training period decreases. However, it was difficult to discern if this trend statistically exists. Therefore, I plotted the forecasts of all eight models against the actual prices of SCIEX to observe all their forecasting accuracies as shown in **Figure 36.**

Table 13. *The regression equation constants and R2 values for the 8 step wise regression models created for SCIEX.*

Training Period (Years)	Equation Constant	Current Price Coefficient	Unemployment Rate Coefficient	CPI Coefficient	Debt Coefficient	Performance Coefficient	Exports Coefficient	R^2
1	138.7	0	0	0	0	-0.3357	-0.992	59%
2	53.57	-0.4926	0	0	0	-0.2574	-0.2752	44%
3	44.55	0	0	0	0.545	-0.2026	-0.3456	47%
4	9.15	0	0	0.1550	0.086	0	0.2986	23%
5	14.00	0	0	0	0.4358	0	-0.0870	35%
6	-33.65	0.1786	2.087	0.2181	0.382	0	-0.2173	37%
7	-34.75	0	1.539	0.2503	0	0	0.1736	33%
8	-3.44	0.2406	1.288	0	0.978	0	-0.1030	35%

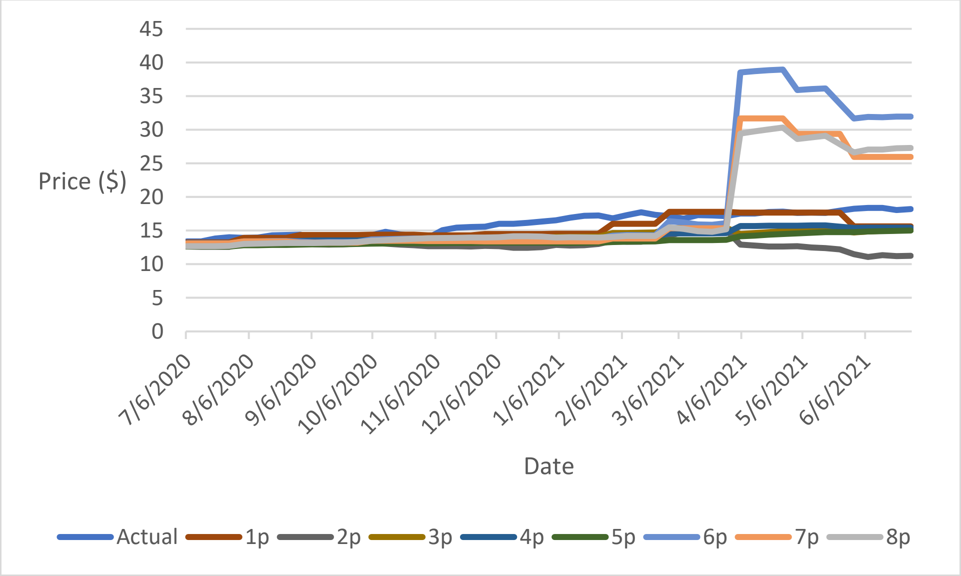


Figure 36. *The forecasted prices for all training period models of SCIEX.*

Like EICPX and BREFX, the dependency of the 6p, 7p and 8p models of SCIEX caused a dramatic increase in their forecasted prices. However, the dependencies of the models dependent on the performance factor were not dramatically affected by the sharp change in the performance factor as discussed. This can be attributed to the performance factor coefficients assigned to them, which were small in comparison to the unemployment rate coefficients assigned to 6p, 7p, and 8p. Because there was such a dramatic change in forecasted price due to the unemployment rate, I decided to swap out the extreme values of the unemployment rate for the forecasted values. Although the 1p, 2p, and 3p forecasts were not dramatically changed by the change in performance factor values, I wanted to be consistent with my previous process of altering the model for these because the forecasts were based on extreme values that the training sets never provided to the models. Furthermore, the 1p, 2p, and 3p models had small training sets comparatively. Therefore, the reaction of these models to the change in performance factor values, albeit small, may be more dramatic than they should have been because they were built on a smaller amount of data and information. To understand how altering the models would affect their prediction accuracy, I decided to first determine the effect of only changing the unemployment rate data values. A plot of these is shown in **Figure 37** below.

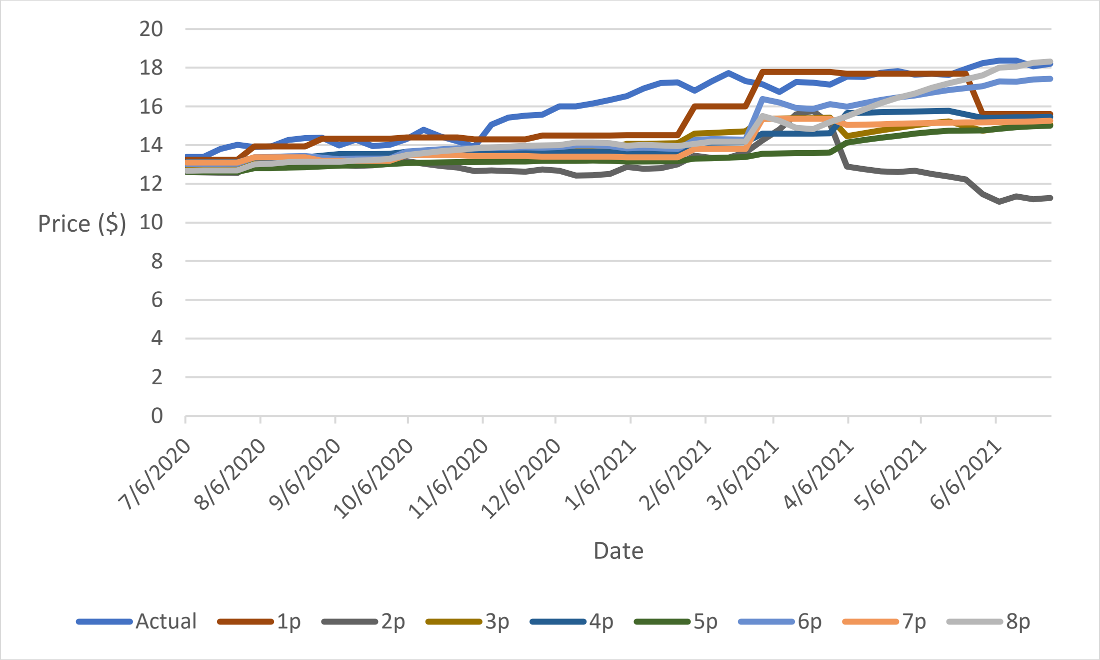


Figure 37. *The forecasted prices for all training period models of SCIEX using altered unemployment rate values.*

In terms of prediction accuracy, it was difficult to determine the superior model from Figure 35 alone. However, the two forecasting models with the lowest MAPE after changing the unemployment rate data were 1p and 6p, which had MAPE's of 5.5% and 8.0% respectively. The plot of just 1p and 6p is shown in **Figure 38**. After altering the unemployment data, I changed both the unemployment rate data and performance factor data together like I did for BREFX. The resulting forecasts are shown in **Figure 39**. Although I created an altered dataset for just the unemployment rate data, I decided not to do so for the performance factor data. if I had done so, the forecasts for 6p, 7p, and 8p would still have been significantly inaccurate.

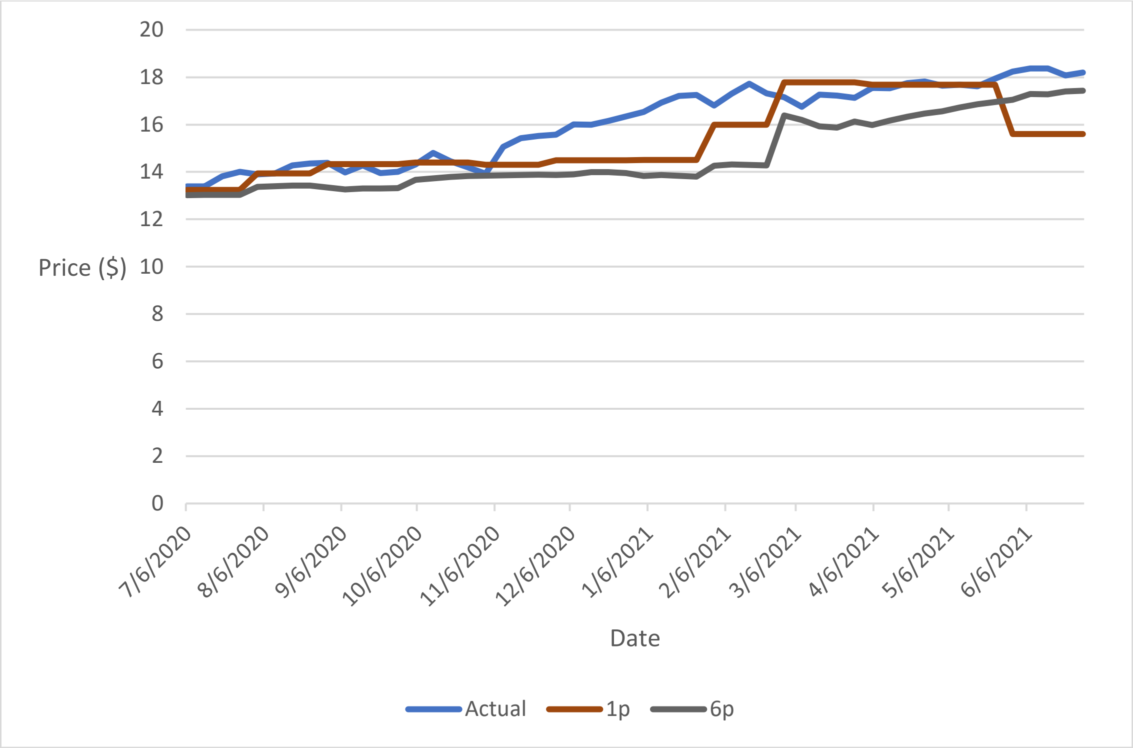


Figure 38. *The forecasted prices of the 6 and 1-year training period models for SCIEX using the altered unemployment rate values.*

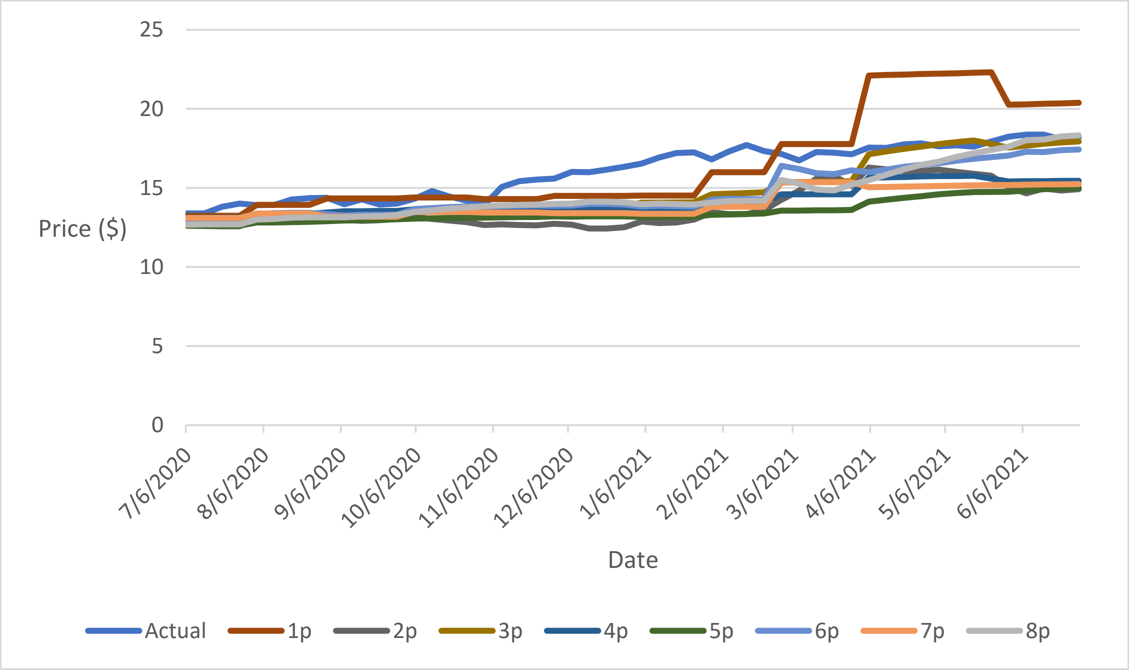


Figure 39. *The forecasted prices for all training period models of SCIEX using the altered unemployment rate and performance index values.*

Like the forecasts in Figure 38, the forecasts in Figure 39 are difficult to differentiate in terms of accuracy due to the number of forecasts plotted. However, in terms of forecasting error, the 3p and 6p models had the lowest MAPE's of 7.4% and 8.0%. While the MAPE of the 3p forecast was reduced compared to the model with only the unemployment data changed, the MAPE of the 6p model did not change. The plots of just these two forecasts are shown in **Figure 40**.

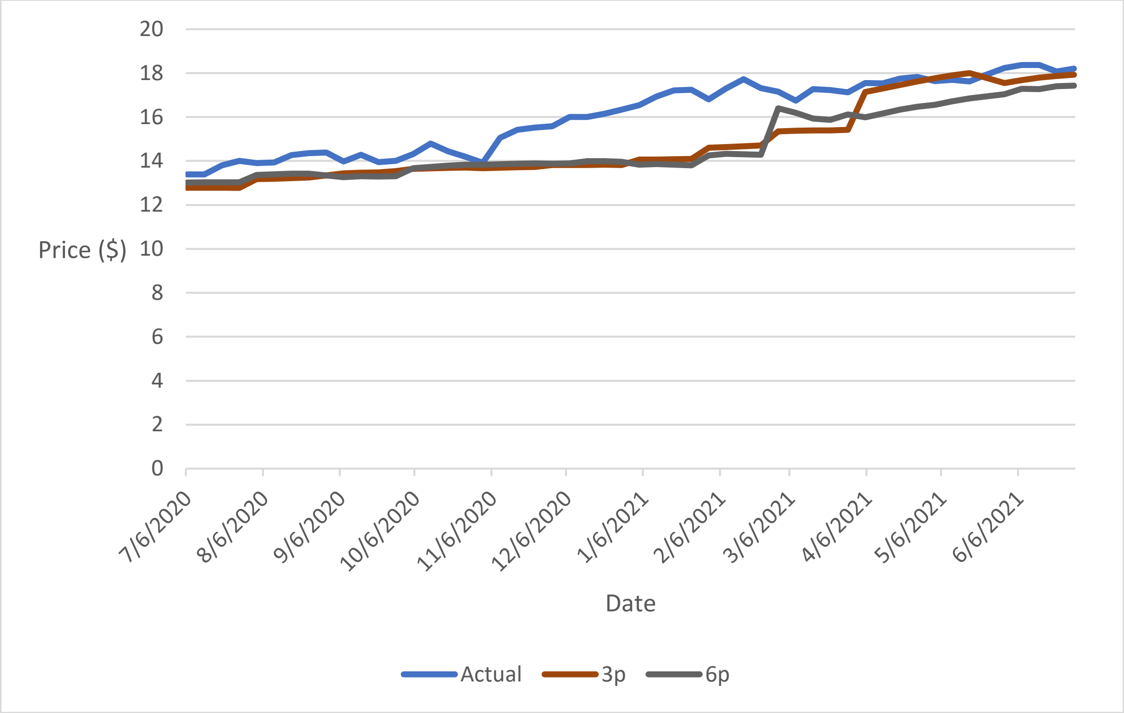


Figure 40. *The forecasted prices of the 6 and 3-year training period models for SCIEX using the altered unemployment rate and performance index values.*

Because I only used one forecasted price per fund for my portfolio optimization, which was the final price of the forecasting period, my inclination was to use the 3p forecast in Figure 40 for the portfolio optimization as it had considerably good accuracy at the end of the period. However, to remain consistent, I chose to use the model that had the overall best forecasting accuracy, not just for the needed prediction point. Therefore, I chose the 1p model for my portfolio optimization because it had an MAPE of 5.5%. However, if I had been using this model to make more forecasts, I would not have chosen the 1p model. As seen with most of the fund models, the eight-year training periods produced the highest accuracy forecasting models. Practically, a one-year training period cannot capture the full range of possible factor values nor the complex relationships of the factors and price response.

References

[1] A. Hayes, “Volatility,” *Investopedia*, 28-Sep-2021. [Online]. Available: https://www.investopedia.com/terms/v/volatility.asp. [Accessed: 01-Oct-2021].

[2] Anonymous "Equity Income Funds Hold Up Well Mutual funds that invest in dividend-paying stocks are a worthwhile conservative choice for part of an equity portfolio," *Postgrad. Med.,* vol. 113, *(3),* pp. 90, 2003. Available: https://www.proquest.com/scholarly-journals/equity-income-funds-hold-up-well-mutual-that/docview/203917994/se-2?accountid=8361.

[3] C. E. Chang and T. M. Krueger D.B.A., "Do Fundamental Index Funds Outperform Traditional Index Funds?" *J. Financ. Plann.,* vol. 28, *(12),* pp. 40-48, 2015. Available: https://www.proquest.com/trade-journals/do-fundamental-index-funds-outperform-traditional/docview/1739060433/se-2?accountid=8361.

[4] C. Spaht and H. Rubin, "Quality Individual Stock Investing Versus Index Investing," *The Journal of Applied Business and Economics,* vol. 18, *(3),* pp. 24-31, 2016. Available: https://www.proquest.com/scholarly-journals/quality-individual-stock-investing-versus-index/docview/1855298142/se-2?accountid=8361.

[5] D. R. Lichtenstein, P. J. Kaufmann and S. Bhagat, "Why consumers choose managed mutual funds over index funds: Hypotheses from consumer behavior," *The Journal of Consumer Affairs,* vol. 33, *(1),* pp. 187-205, 1999. Available: https://www.proquest.com/scholarly-journals/why-consumers-choose-managed-mutual-funds-over/docview/195898554/se-2?accountid=8361. DOI: http://dx.doi.org/10.1111/j.1745-6606.1999.tb00766.x.

[6] G. Iacurci, “Some mutual funds are pricier than others. here's when they may benefit investors,” *CNBC*, 24-Nov-2020. [Online]. Available: https://www.cnbc.com/2020/11/24/heres-when-active-mutual-funds-tend-to-outperform-index-funds.html. [Accessed: 08-Sep-2021].

[7] J. Chen, “Institutional shares,” *Investopedia*, 27-Sep-2021. [Online]. Available: https://www.investopedia.com/terms/i/institutionalshares.asp. [Accessed: 28-Mar-2022].

[8] J. Fernando, “What the price-to-book ratio (P/B ratio) tells you?,” *Investopedia*, 14-Sep-2021. [Online]. Available: https://www.investopedia.com/terms/p/price-to-bookratio.asp. [Accessed: 01-Oct-2021].

[9] J. K. Glassman, “The Truth About Index Funds,” *Kiplinger's Personal Finance*, pp. 31–32, Oct-2021.

[10] Kopp, C. M. (2021, September 4). *Strategies to maximize your 401(k)*. Investopedia.https://www.investopedia.com/articles/personal-finance/091515/best-strategies-maximize-your-401k.asp.

[11] "Morningstar, inc..," *Morningstar*. [Online]. Available: https://www.morningstar.com/funds. [Accessed: 08-Oct-2021].

[12] Parker, K., & Fry, R. (2020, July 27). *More than half of U.S. households have some investment in the stock market*. Pew Research Center. https://www.pewresearch.org/fact-tank/2020/03/25/more-than-half-of-u-s-households-have-some-investment-in-the-stock-market/.

[13] W. Mun Fong and Z. Ong, "The Long and Short of Profitable Dividend Yield Strategies," *The Journal of Wealth Management,* vol. 18, *(4),* pp. 124-137, 2016. Available: https://www.proquest.com/scholarly-journals/long-short-profitable-dividend-yield-strategies/docview/1932133267/se-2?accountid=8361. DOI: http://dx.doi.org/10.3905/jwm.2016.18.4.124.