University of Michigan
School of Information
SIADS 696 - Milestone II

Project Title
S&P 500 Algorithmic Trading Bot

Team Members
Calvin Raab
Mijee Kim
Youngho Shin

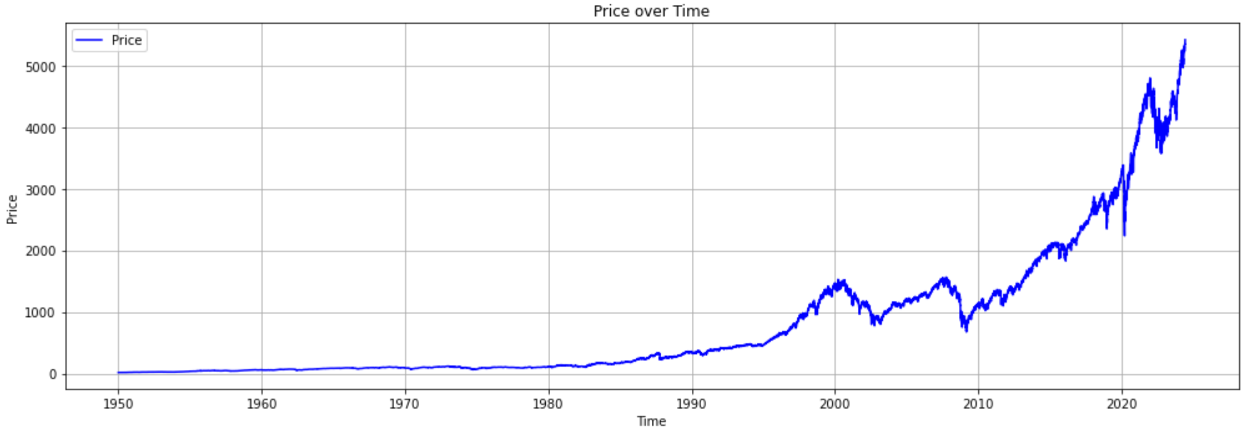
Advisor
Noha Ghannam

Github Repository
https://github.com/calvinraab/ML_Algorithmic_Trading

Date
June 27, 2024

Introduction

This paper details the makings of a Machine Learning algorithm to predict whether or not you should, buy, hold, or sell the S&P 500 index. The S&P 500, Standard and Poors 500 Index, is a stock market index that measures the performance of 500 of the largest publicly traded companies in the United States of America. By leveraging both supervised and unsupervised learning techniques, we aim to create a model that can accurately predict market behavior and assist in making profitable trading decisions. A successful model would offer a powerful tool to enhance trading strategies, potentially leading to better financial outcomes. Additionally, it would contribute to both academic and practical understanding of the factors influencing stock market movements, thereby aiding in the development of more sophisticated financial models. The motivation for working on this project stems from the historical performance of the S&P 500 index, which has shown a consistent increase with an 8% compound annual growth rate (CAGR) from 1950 to 2023. This performance significantly outpaces the average inflation rate in the US of about 3.5% over the same period[1], establishing the S&P 500 as a stable investment option capable of beating inflation. Our goal is to leverage machine learning to generate returns slightly higher than this historical CAGR, enhancing the investment strategy and maximizing financial outcome.



Along the way in this paper the team has to make many assumptions that we will discuss. The team understands that predicting the market has been an effort taken on by many and if there truly was a simple way to do it, that would be done. There have even been papers that in a perfect market scenario the stock market movements reflect a random walk[2]. Although, what motivates us is this paper is research that the stock market does display some short term predictability[3]. The team hopes to discover some of these trends while building a profitable trading bot.

We will be using both supervised and unsupervised learning algorithms during the creation of our algorithmic trading bot. In fact, our unsupervised learning bot will actually give us a parameter that will then be passed into our supervised learning bot. For the unsupervised portion we will be using a k-means algorithm to group together economic data in order to find groups of economic "moods", grouping together good, bad and average economies. That will then naturally lead into supervised learning where we will be training multiple supervised learning models to predict whether or not one should, buy, sell, or hold S&P 500 stock.

The main finding during our supervised learning portion was that we could build a model that got over a 50% accuracy on novel data in a cross evaluation training method. We were able to find two different supervised learning methods that had over 50% accuracy. Through further analysis of the parameters we also uncovered that the most important feature when building and predicting this model was the unemployment rate difference year over year.

Related Work

Our first example of related work is a paper titled “Stock Market Prediction with High Accuracy using Machine Learning Techniques” by Malti Bansal, Apoorva Goyal, and Apporva Chaudhary[4]. This paper is doing the same prediction task as we plan to do, building a predictive stock market model. In their work they use prior price action and include multiple models in their work: K-Nearest Neighbors, Linear Regression, Support Vector Machine, Decision Tree Regression, and Long Short-Term Memory. Our approach will look similar to this project with some fundamental differences. The key difference that differentiates our project is that we are not just using price action alone but we will also be including important economic factors into our analysis to try and capture the price action tied to economic factors.

Another similar project is the use of machine learning in stock market prediction as explored in various studies, such as those found on sites like Kaggle[5]. For example, [Tesla Stock Price Prediction using LSTM]* uses similar datasets to predict stock movements. This project primarily uses supervised learning with financial indicators as features. Our project differentiates by incorporating unsupervised learning to uncover underlying patterns and correlations not immediately apparent, aiming to improve prediction accuracy and adaptability.

The paper "Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions" by Nusrat Rouf and others[6] reviews various machine learning methods for predicting stock market prediction. It highlights techniques such as SVM, ANN, DNN, and hybrid approaches, combining market data with textual data from social media a `nd financial news to enhance prediction accuracy. This project, like our project, uses machine learning techniques for stock market prediction and emphasizes the inclusion of various data, such as economic indicators, to predict prices. However, while this project explores various methodologies, our goal is to develop a trading tool that generates actionable signals specifically for the S&P 500.

Data Sources

S&P 500 Price Action[7]

This dataset includes daily, monthly, and annual price information for the S&P 500 index, dating back to 1927. It is crucial for understanding the long-term trends, volatility, and patterns in the stock market. This data forms the backbone of the supervised learning model, allowing the prediction of future price movements based on historical trends.

Real Gross Domestic Product (GDP)[8]

This data is broken down by quarter and dates back to Q1 1947. Real GDP measures the value of economic output adjusted for price changes (inflation or deflation). It is a key indicator of the economic health and growth of the country, providing context for stock market performance as economic expansion typically leads to higher corporate earnings and stock prices.

Unemployment Rate[9]

This dataset provides monthly unemployment rates dating back to 1947. The unemployment rate is a measure of the percentage of the labor force that is jobless and actively seeking employment. High unemployment rates can indicate economic distress, which can negatively impact stock market performance.

Consumer Price Index (CPI)[10]

The CPI measures the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services. It is a primary indicator of inflation. Understanding inflation trends is vital because high inflation can erode purchasing power and affect corporate profitability, thus impacting stock prices.

Federal Funds Rate[11]

This data, dating back to 1954, tracks the Federal Funds Rate. The federal funds rate influences overall economic activity, including consumer spending and business investment. Changes in this rate can significantly impact stock market performance by affecting borrowing costs and investment returns.

Feature Engineering - Data Processing

We decided to conduct our analysis on an annual basis since not all data was available daily. To create a comprehensive dataset, we averaged the data over each year, resulting in a 1-year timeframe dataset.

As can be seen in the Github repository, extensive data cleaning was necessary in order to get all of the data to have the ability to get merged with each other. For the year datasets, if the economic indicator did not have a January first date, I took the average of all the data for that year. The year over year percentage data that we use, there are varying ways that that value is calculated.

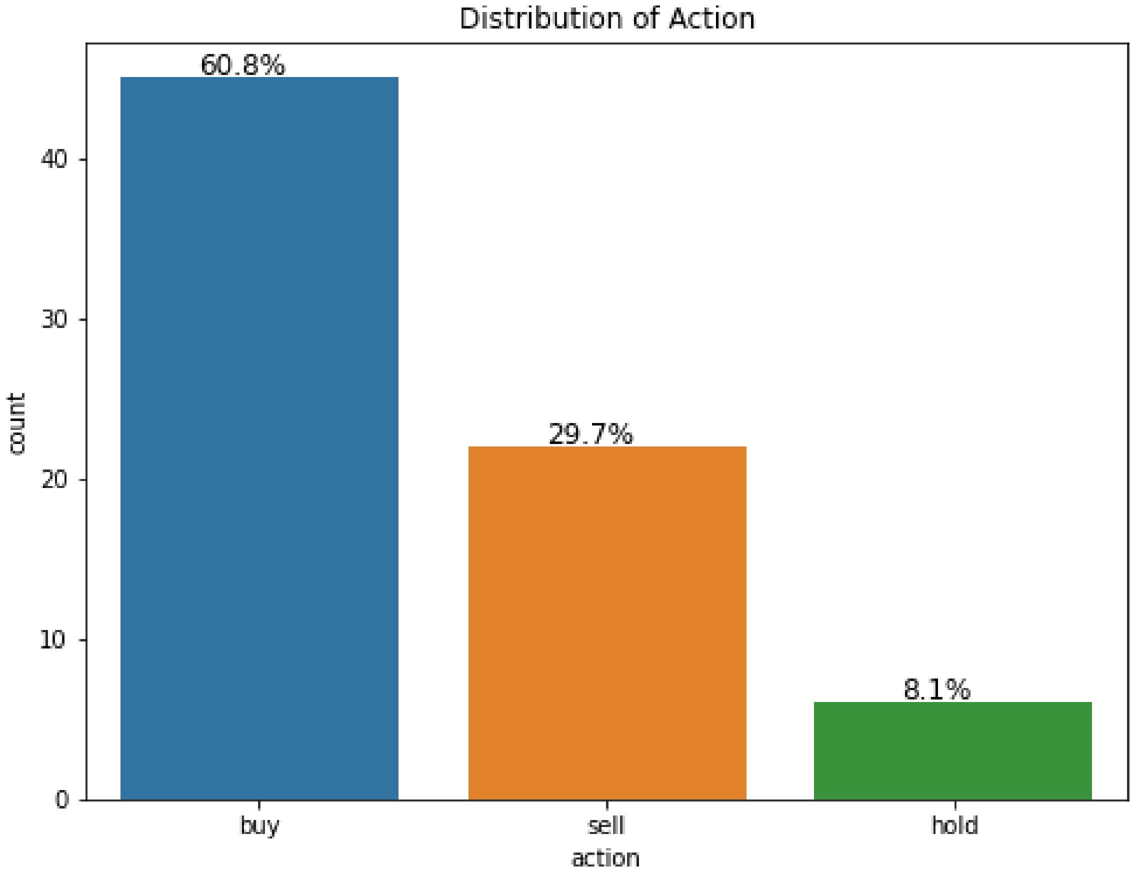
One of the critical decisions we made involved defining the dependent variable—our action signal—based on the year-over-year percentage change in the S&P 500 index price. We established a 10% threshold for a buy signal to target a 10% annual return, which is approximately 30% higher than the S&P 500 index's compound annual growth rate (CAGR) of about 8% from 1950 to 2023.

The specific criteria we used to assign these signals were:

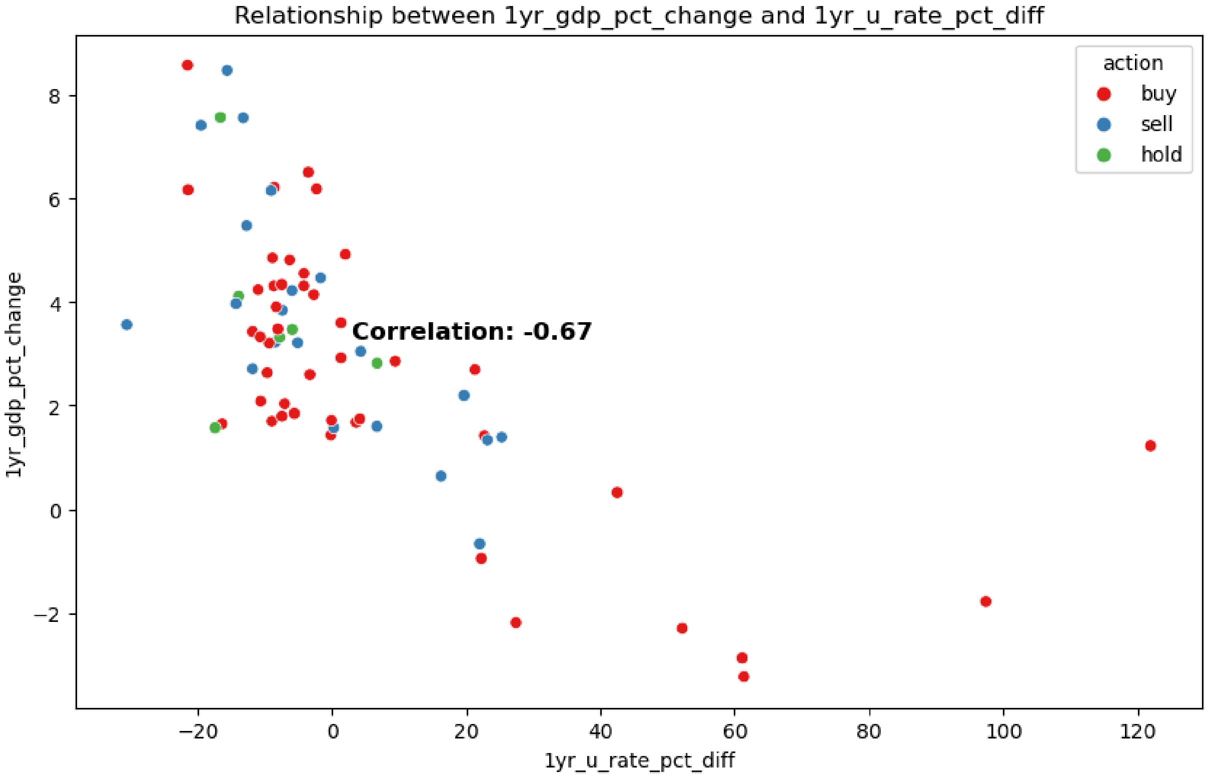
1. **Buy**: If the next year's price year-over-year percentage change (price_yoy_pct_change) is over 10%.

2. **Hold**: If the next year's price_yoy_pct_change is between 5% and 10%.
3. **Sell**: If the next year's price_yoy_pct_change is under 5%.

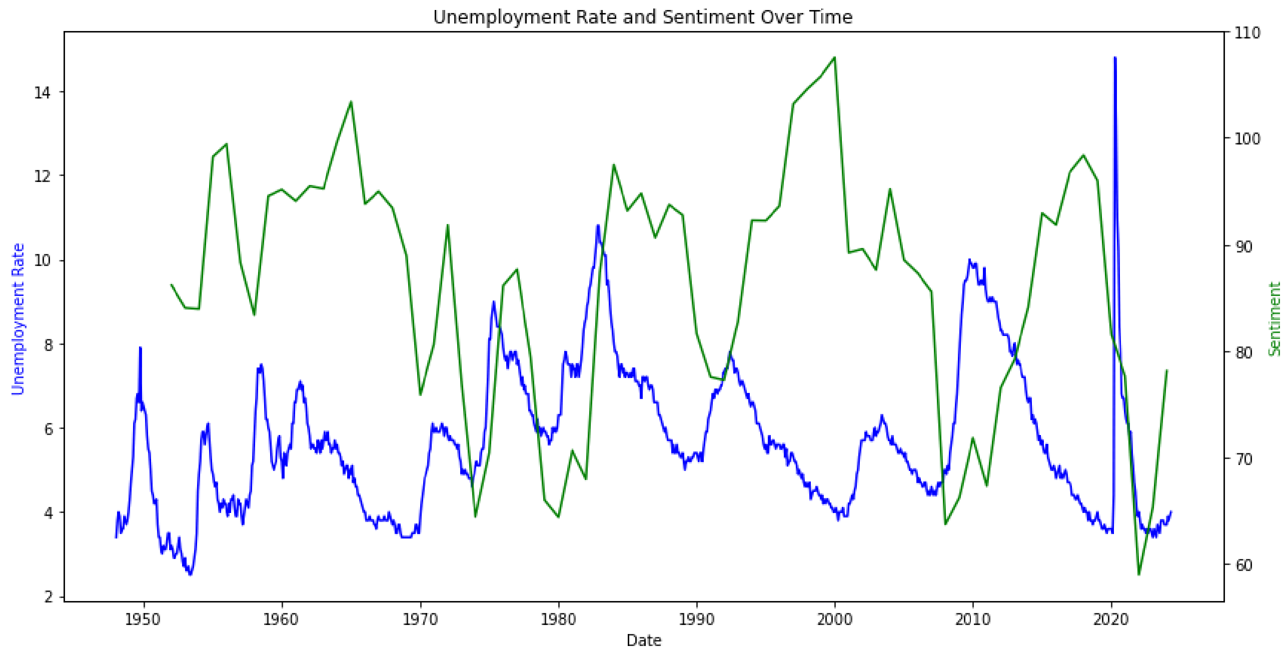
The reason for pinning this to out do even the CAGR is to give us some margin for error. Originally we thought about doing something where if the market loses money, then it would be a sell. No change or 1% change in either direction, hold. And if positive then buy. Although logically this makes sense as there is such large volatility in stock price, we wanted to give ourselves more margin for error.



The predominance of buy signals (60.8%) suggests that, historically, the S&P 500 has frequently experienced periods of strong performance, with annual returns exceeding 10% more often than not. This pattern reflects the long-term upward trend and overall growth trajectory of the market. Conversely, the low percentage of hold signals (8.1%) implies that the market seldom experienced moderate growth periods (5-10% returns).



The scatter plot depicts the relationship between the 1-year unemployment rate percentage difference (1yr_u_rate_pct_diff) and the 1-year GDP percentage change (1yr_gdp_pct_change). There is a significant negative correlation (-0.67), indicating that as the unemployment rate decreases, the GDP tends to increase, and vice versa. Particularly noteworthy is that in cases where GDP change is low or negative, and the unemployment rate change is high, the "BUY" action is predominantly observed.This pattern highlights the inverse relationship between unemployment and economic growth, influencing investment actions accordingly.



The unemployment rate and sentiment both exhibit periodic fluctuations with distinct highs and lows. There are instances where peaks in the unemployment rate (high unemployment) correspond with troughs in sentiment (low sentiment), suggesting an inverse relationship between these two variables. Significant spikes or drops in both the unemployment rate and sentiment likely correspond with major economic events, such as recessions, financial crises, or recovery periods.

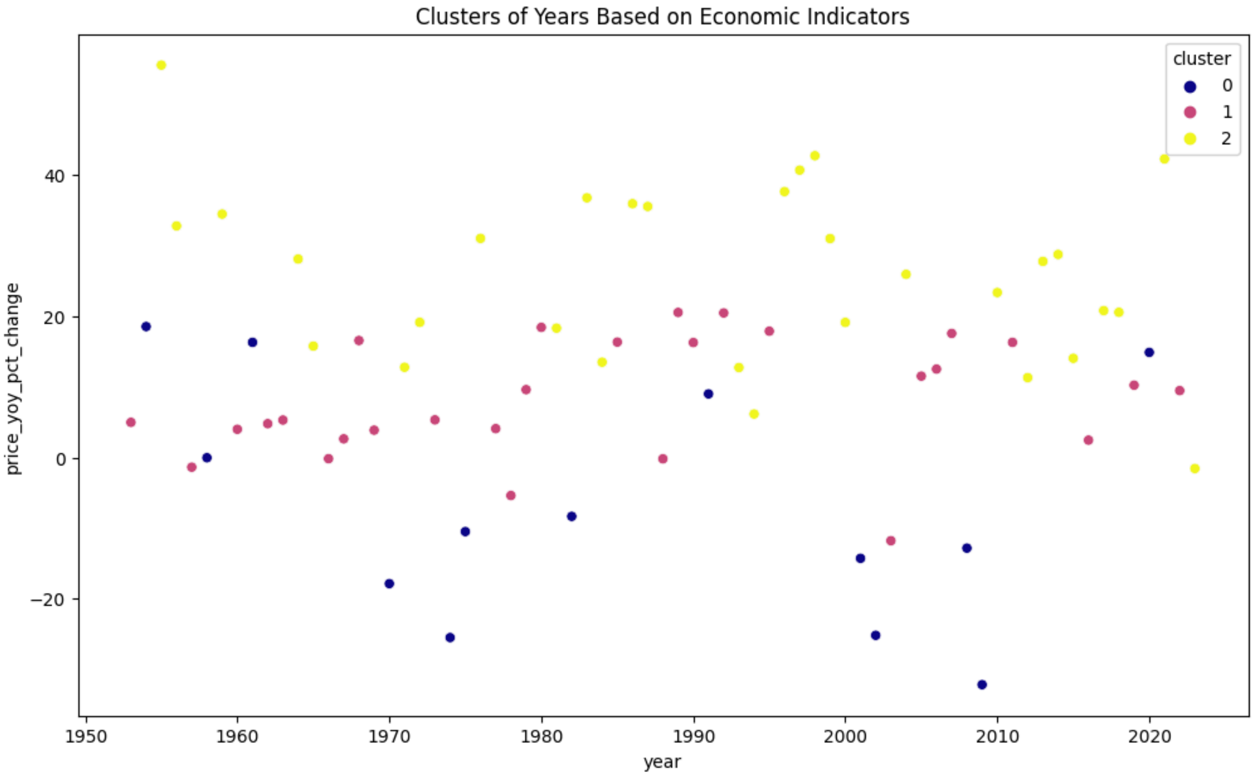
[Final features]

Unsupervised learning	price_yoy_pct_change
	1yr_u_rate_pct_diff
	1yr_sentiment_pct_change
	1yr_gdp_pct_change
Supervised learning	price_yoy_pct_change
	volume_yoy_pct_change
	1yr_u_rate_pct_diff
	1yr_sentiment_pct_change
	1yr_gdp_pct_change
	economy_label_encoded

Unsupervised Learning

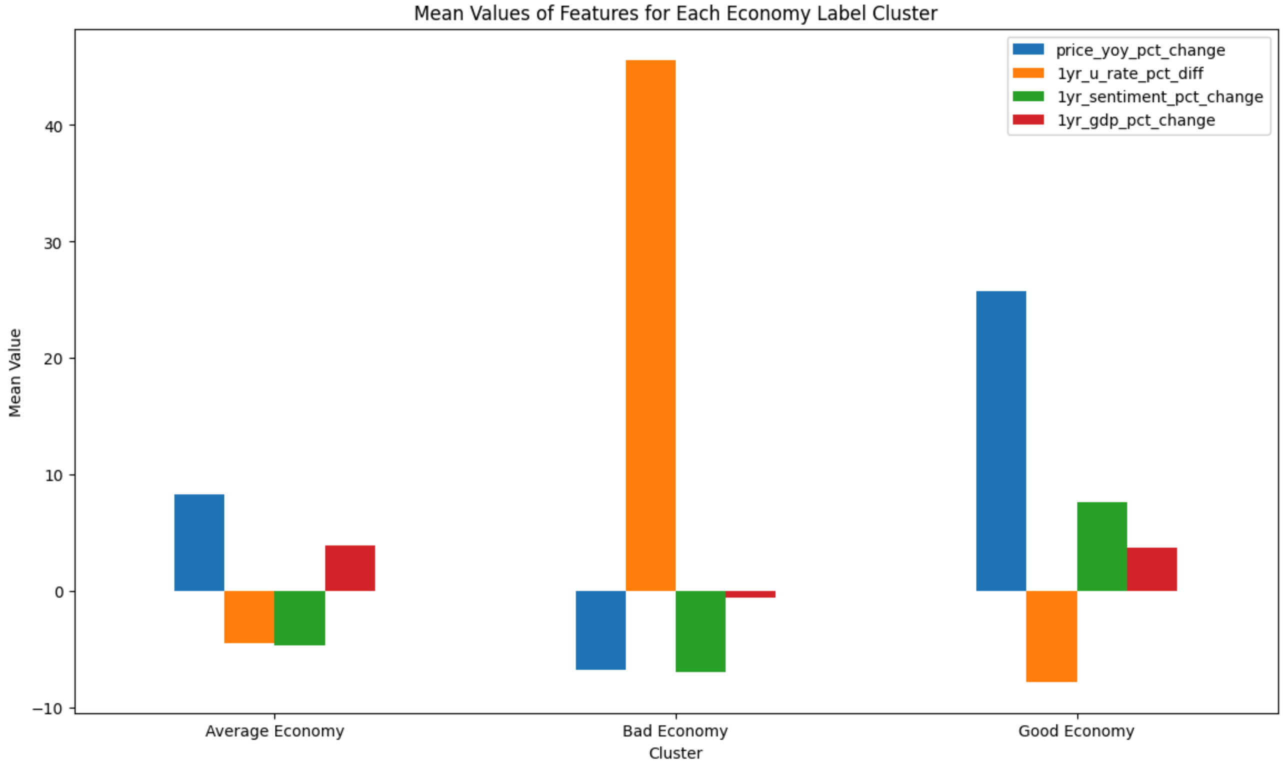
We began by preparing our dataset, which included key economic indicators such as year-over-year percentage changes in price and volume, one-year percentage differences in the unemployment rate, one-year sentiment percentage changes, and one-year GDP percentage changes. We ensured the dataset was complete by removing any rows with missing values. Next, we standardized the features using the StandardScaler. Standardization was crucial to ensure that all features contributed equally to the clustering process, as they were on different scales. We selected key features—price year-over-year percentage change, one-year unemployment rate percentage difference, one-year sentiment percentage change, and one-year GDP percentage change for clustering analysis.

We then applied the K-Means clustering algorithm, specifying three clusters to categorize the data into distinct groups. The K-Means algorithm assigns each data point to one of the three clusters based on similarities in the selected features. The number of clusters was chosen based on our hypothesis about the potential economic states, although this number could be adjusted based on further analysis.

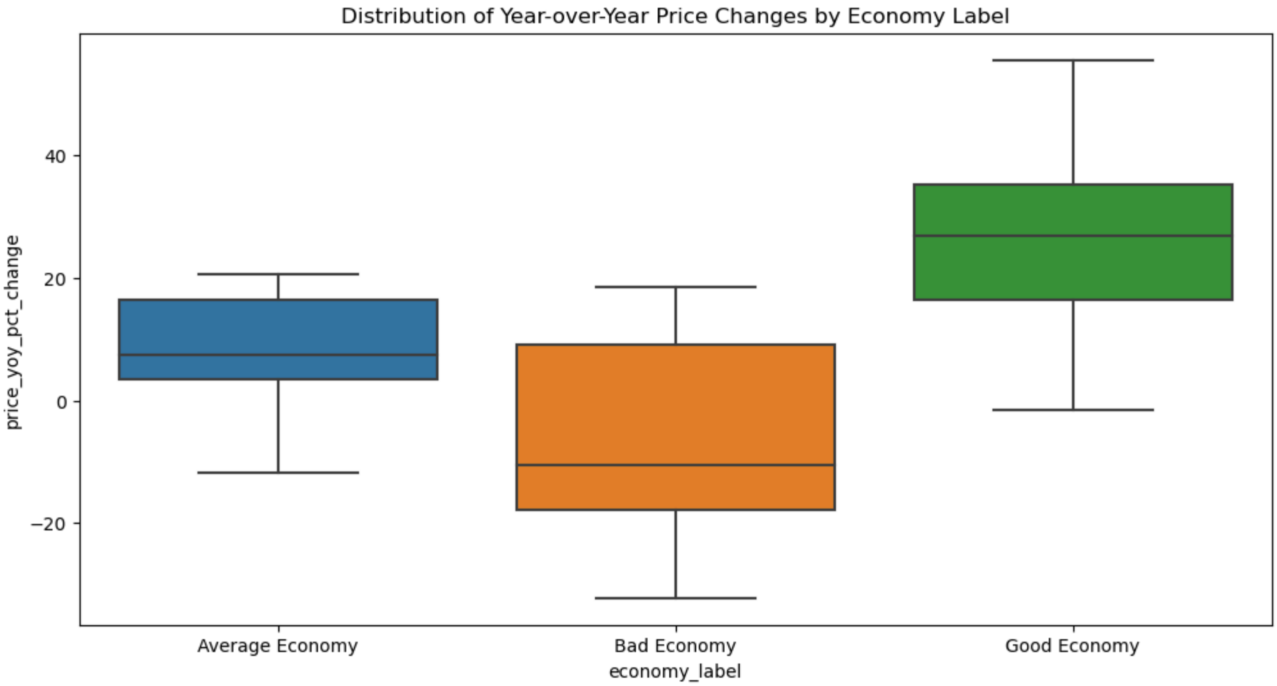


We created a scatter plot with the year on the x-axis and the price year-over-year percentage change on the y-axis. The clusters were represented by different colors, allowing us to observe how different years grouped together based on their economic indicators. The visualization revealed three distinct clusters, each representing different economic states. Cluster 0 appears to be a bad economy. Cluster 1 appears to be an average economy and cluster 2 appears to be a good economy.

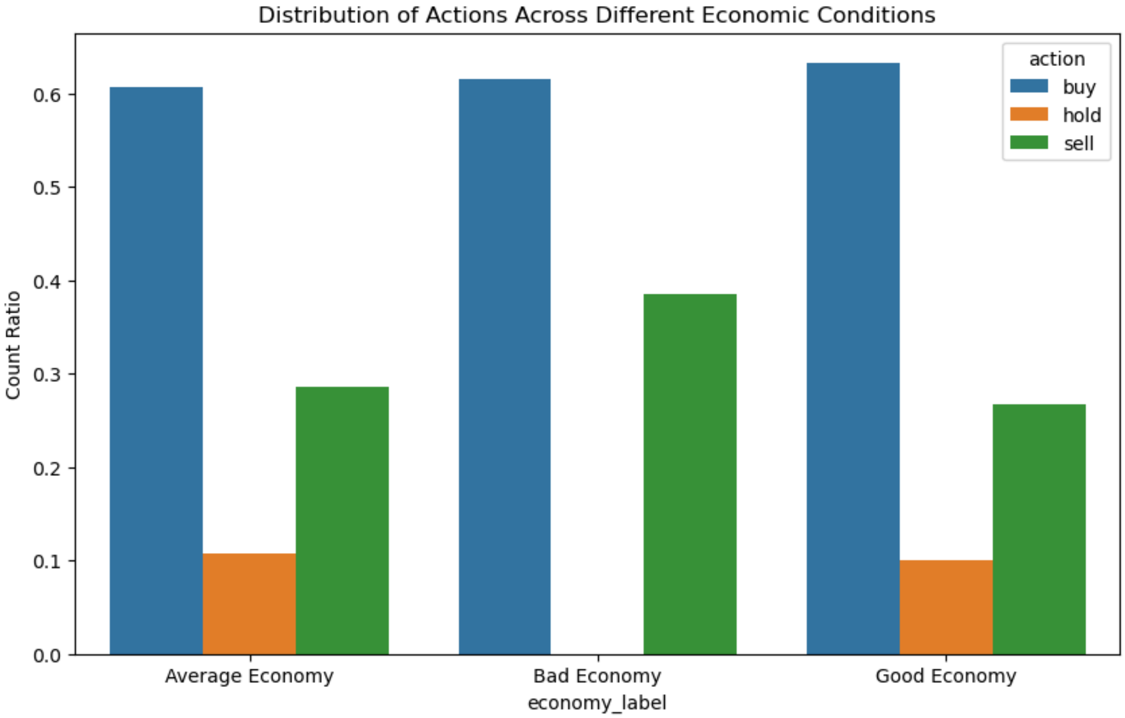
It is important to realize that despite these clusters looking very good, there are failings that come with the K-Means algorithm. Firstly, we can see that the groupings are not perfect. There is a lot of bleeding between clusters. Cluster number 1, our “average economy” has values interlaid with both other clusters. There are other failings of using K-means such as there having to be a fixed number of clusters and the assumption that clusters will be roughly the same size.



The above visualization is helpful because it helps to see how the clusters look plotted against our indicator of note, S&P 500 price change by percent. However, these clusters are created using all of the economics indicators we fed in, not just the percent price change. To confirm the hypothesis that these were solid moods of the economy, it is important to see them graphed out to see all the economic indicators as well. In looking at each of the clusters broken down by month our hypothesis that the cluster seemed to capture moods in the stock market is confirmed. Cluster 0 has all of the makings of a very bad economy.



The visualization shows the price change across different economic labels. In a Good Economy, price changes are positive, while in a Bad Economy, price changes are negative. An Average Economy falls in between. This indicates that the clustering process was well-executed.



The visualization reveals that volume changes are more volatile in a Bad Economy. In fact, the distribution of action labels shows that during a Bad Economy, actions are primarily composed of Buy and Sell.It is interesting to see in the visual that in a bad economy, a hold signal never showed up.

Supervised Learning

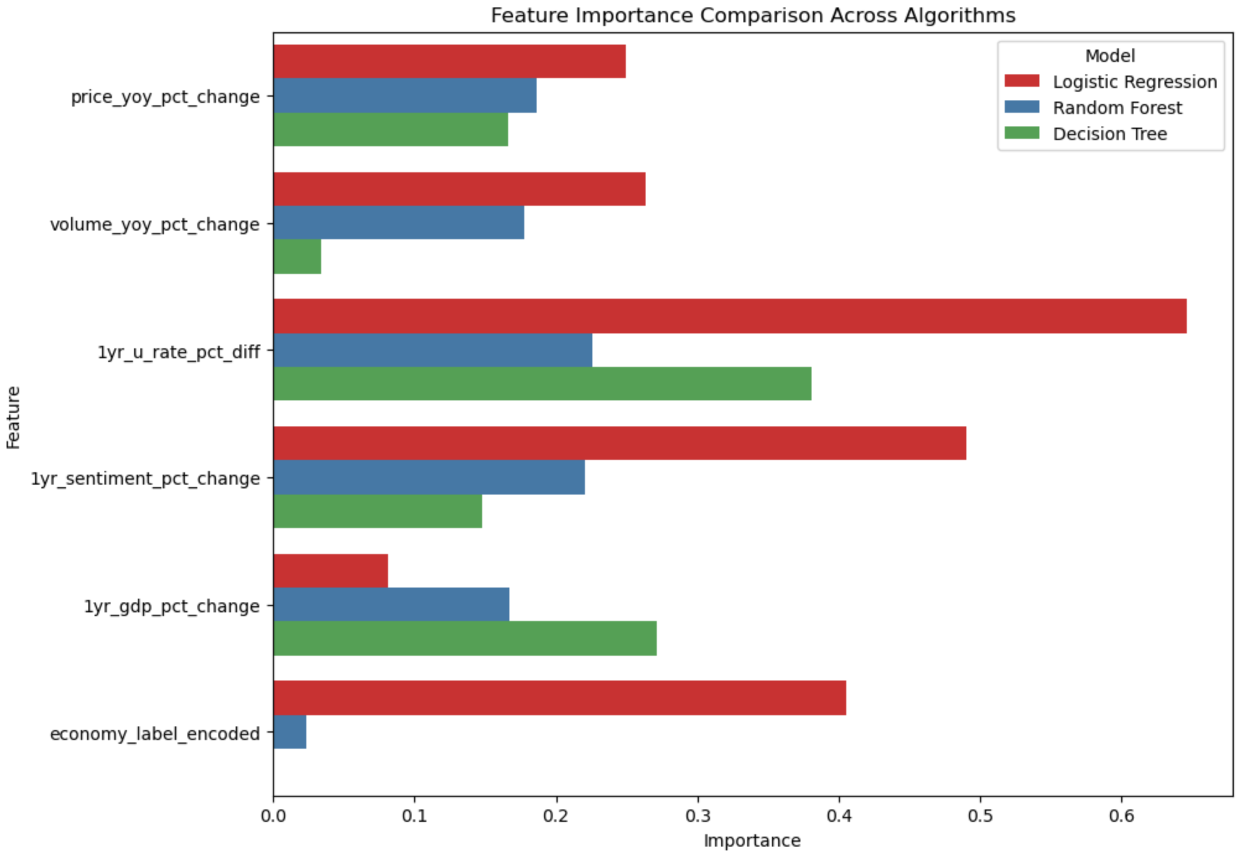
To preprocess the data for supervised learning, we one-hot encoded the categorical variables. We selected features for the model, including several percentage changes and the encoded economy label, with the target variable set as the encoded action. The data was split into training and testing sets, and the features were standardized. This is a classification problem, trying to build a classifier for the three actions we have here.

Firstly, a Logistic Regression model is initialized, which is a linear classifier that estimates probabilities using a logistic function. Secondly, a Decision Tree Classifier is employed, which constructs a tree structure by splitting the data based on feature values, aiming to maximize information gain in each leaf node. Decision trees are capable of capturing complex interactions between features. Lastly, a Random Forest Classifier is utilized, which is a method consisting of multiple decision trees trained on different samples of the data and using random subsets of features. Each model undergoes 5-fold cross-validation to estimate its accuracy and variance.

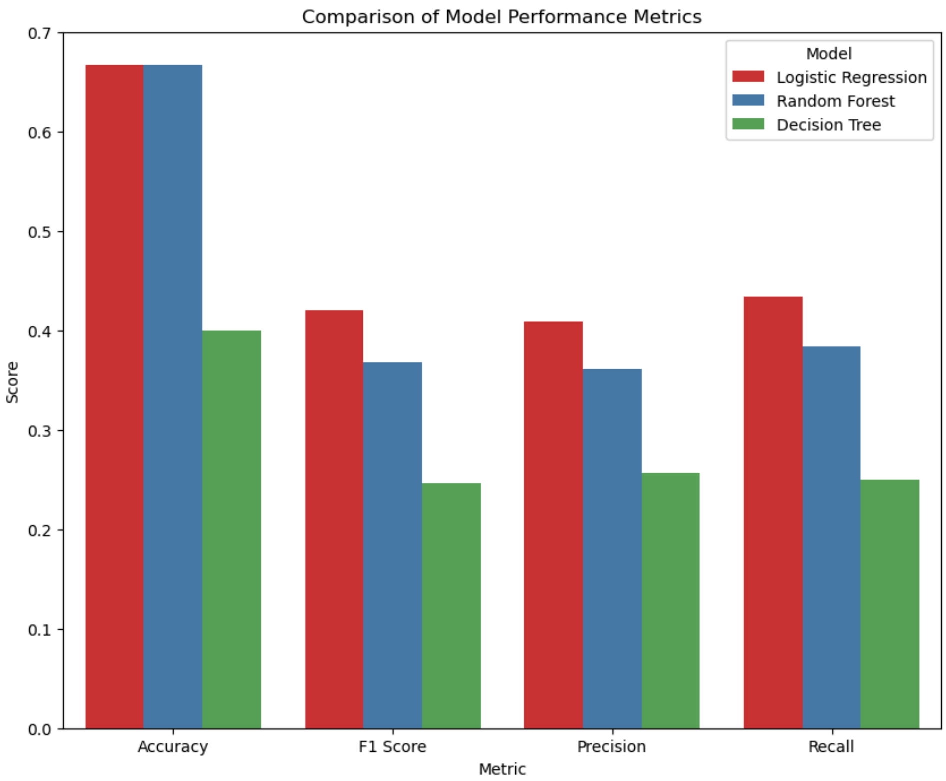
Even with the unsupervised learning having been as successful as it was, the team was still questioning how well the supervised learning, actually predictive elements, would work. The categorization worked but we were still unsure if any of these columns had actual predictive elements. To our surprise, the supervised learning models worked much better than expected. We used Accuracy as the score in which we would be most focused on. Our logistic regression model when tested with novel data has a 67% accuracy. Originally we were shooting for an accuracy of over 50%, a 67% accuracy we were very happy with. We didn’t just run a logistic regression model, we wanted to try multiple supervised learning models to see if different models performed better. Next we ran a decision tree. This model did not perform as well and had a 40% accuracy. Next we ran a random forest classifier. This model matched the performance of the logistic regression model with a 67% accuracy.

Despite the success of two of our models, these supervised models have their own set of failings. These models have many of the same failings. Each of them struggle with the bias-variance tradeoff as they try not not to overfit or underfit. These models can also overfit the data and are sensitive to outliers. The fact that we don’t have that much data we are pumping into these models also is a pain-point for us and the reason we had to do the 5 fold cross validation in order to try and build the best model possible. For a deeper understanding of our failure analysis I will look through the 5 incorrectly classified values. Most of these are instances where our model incorrectly predicted a sell vs hold. It is hard to say exactly why these were incorrectly classified but most of them have very high year over year percent change which could have contributed.

We wanted to further explore these models and pull more information out of them. We went through and evaluated the feature importance on each of the features of the dataset. A very interesting thing we found was that the percent change in unemployment rate was for each model the most important feature. It is interesting to think that unemployment can be predictive for the S&P 500 price. Another interesting takeaway is that the models all did not allocate the same feature importance to the same features. For instance in the Decision Tree model GDP percent change was the second most important feature and in the Random Forest model GDP percent change was second to last.



The visualization displays feature importance across three algorithms: Logistic Regression, Decision Tree, and Random Forest. The most significant feature, "1yr_u_rate_pct_diff," is crucial in predicting the target variable across all models. This feature's high importance suggests that the one-year percentage difference in the unemployment rate strongly influences the models' decisions. This emphasizes the need to focus on unemployment rate changes for accurate predictions.



The visualization displays the performance metrics of three models: Logistic Regression, Decision Tree, and Random Forest. For each model, accuracy, precision, recall, and F1 score are calculated and combined into a single dataframe, which is then visualized using a bar plot. This approach allows for a clear comparison of the performance of the three models across multiple metrics. From the graph, it is evident that Logistic Regression and Random Forest outperform the Decision Tree in all metrics. Logistic Regression shows high accuracy, recall, and F1 score, indicating balanced performance. Random Forest also demonstrates high accuracy and precision, making it another strong performer. Decision Tree, however, lags behind in all metrics, showing particularly low precision and recall. This visualization effectively highlights which model performs best in each metric, facilitating a clear understanding of the strengths and weaknesses of each model.

Discussion

As discussed throughout the paper there was a lot that the team learned and were surprised by throughout the process. When it comes to our unsupervised process, the biggest thing for us was the fact that the clusters were able to be interpreted. We were concerned that this data could not actually be grouped into economic moods, but we were pleasantly surprised how interpretable the clusters were. We struggled a bit with trying to explain why some of the groups were the way they were. When it comes to supervised learning we were surprised by the fact that we could build a supervised model that actually made accurate predictions over 50%. We were unsure if any of the columns would actually have any predictive power. We struggled with a lot of the model interpretability as a lot of the coefficients the model chose were hard to interpret.

Ethical Considerations

Generally, the data we use does not contain unethical information. However, despite using reliable sources like FRED and Yahoo, data collection errors can occur, leading to inaccuracies. These inaccuracies may result in incorrect conclusions. Therefore, it is essential to provide clear, limited interpretations of our findings to prevent readers from making misguided decisions. By highlighting the potential limitations and provisional nature of our results, we ensure that users approach the insights with caution and seek further validation before acting on them. It is important not to pass this off as a money making algorithm as there is so much variance in the stock market and we made a lot of assumptions.

Future Work

There is a lot of future work that can be built from this. Firstly a further analysis of our models should be done. More explainability in why we got the model we have is important. Another avenue of future work would be running our model on other stocks and indexes. Maybe on some that didn't go up so much over the span of the data. Seeing how these models compare with the S&P model we have here we feel would be very beneficial and interesting. Another idea for future work that I think would be very interesting would be throwing more complex models at this. Putting this financial data through a neural network would make for an interesting project as a neural network would be able to discover more relationships in the data than the supervised learning models that we used here.

Statement of Work

Calvin contributed to all parts of the data science lifecycle. Calvin set up the initial repository. Calvin brought all the datasets into python and processed all the data, preparing it for modeling. Calvin shared this dataset with the team for visualization. Calvin then built out the unsupervised and supervised machine learning models and analyzed their effectiveness in building a machine learning bot.

Mijee played a crucial role in the exploratory data analysis (EDA) by contributing to the visualization efforts and setting reasonable thresholds for the target variable. Mijee also interpreted various visualizations and modeling results, ensuring that the data was effectively analyzed and the insights were clearly communicated.

Youngho was responsible for data collection, preprocessing, and visualization. Youngho explored various datasets from FRED and selected useful datasets for the team. Additionally, Youngho conducted extensive data visualizations during the exploratory data analysis (EDA) phase. After the model results were obtained, he contributed to interpreting the results through various visualization techniques.

References

1. U.S. Inflation Rate by Year: 1929 to 2024, Investopedia, https://www.investopedia.com/inflation-rate-by-year-7253832, June 2024
2. Lo W Andrew. "Efficient Market Hypothesis." Massachusetts Institute of Technology, https://web.mit.edu/Alo/www/Papers/EMH_Final.pdf#:~:text=URL%3A%20https%3A%2F%2Fweb.mit.edu%2FAlo%2Fwww%2FPapers%2FEMH_Final.pdf%0AVisible%3A%200%25%20. Accessed June 2024.
3. Malkiel G Burton. "The Efficient Market Hypothesis and Its Critics." University of California Berkeley, https://eml.berkeley.edu/~craine/EconH195/Fall_14/webpage/Malkiel_Efficient%20Mkts.pdf#:~:text=URL%3A%20https%3A%2F%2Feml.berkeley.edu%2F~craine%2FEconH195%2FFall_14%2Fwebpage%2FMalkiel_Efficient%2520Mkts.pdf%0AVisible%3A%200%25%20. Accessed June 2024.
4. Bansal Malti, Apoorva Goyal, Apoorva Choudhary, “Stock Market Predicition with High Accuracy using Machine Learning Techniques”, https://www.sciencedirect.com/science/article/pii/S1877050922020993 Accessed June 2024.
5. Goel Shwetanshu, “Tesla Stock Price Prediction using LSTM, https://www.kaggle.com/code/majinx/tesla-stock-price-prediction-using-lstm Accessed June 2024.
6. Rouf, Nusrat, Majid Bashir Malik, Tasleem Arif, Sparsh Sharma, Saurabh Singh, Satyabrata Aich, and Hee-Cheol Kim. 2021. "Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions" *Electronics* 10, no. 21: 2717. https://doi.org/10.3390/electronics10212717
7. S&P 500 historical data, Yahoo Finance, https://finance.yahoo.com/quote/%5ESPX/history/?period1=-1325583000&period2=1718232014
8. "Real Gross Domestic Product." Federal Reserve Bank of St. Louis, Federal Reserve Economic Data, https://fred.stlouisfed.org/series/GDPC1. Accessed June 2024.
9. "Unemployment Rate." Federal Reserve Bank of St. Louis, Federal Reserve Economic Data, https://fred.stlouisfed.org/series/UNRATE. Accessed June 2024.
10. "Inflation, Consumer Prices for the United States." Federal Reserve Bank of St. Louis, Federal Reserve Economic Data, https://fred.stlouisfed.org/series/FPCPITOTLZGUSA. Accessed June 2024.
11. "Effective Federal Funds Rate." Federal Reserve Bank of St. Louis, Federal Reserve Economic Data, https://fred.stlouisfed.org/series/DFF. Accessed June 2024.