DATA602 Final Project Report

- **Course Code and Name:** DATA602 - Principles of Data Science
- **Semester and Year:** Fall 2024
- **Instructor Name:** Dr. Fardina Alam
- **Group Project Name:** Analyzing S&P 500 Performance and U.S. Presidential Election Outcomes
- **URL to Final Tutorial:** https://reetikaagajula2002.github.io/Analyzing-S-P-500-Performance-and-U.S.-Presidential-Election-Outcomes/s&p500_election_predictor_analysis.html
- **Group Members:** Reetikaa Gajula, Mark Jeakle, and Sylvia Miller
- **Date of Submission:** 15 December 2024

Contributions

- **Project idea:** Equal Contribution - every member found a project idea and did data preprocessing and visualizations on the dataset before the entire group chose the S&P 500 and U.S. Presidential Election dataset presented by Mark Jeakle
- **Dataset Curation and Preprocessing:** Predominantly Mark Jeakle with help from Reetikaa Gajula and Sylvia Miller
- **Data Exploration and Summary Statistics:** Equal Contribution
- **ML Algorithm Design/Development:** Predominantly Reetikaa Gajula with help from Mark Jeakle and Sylvia Miller
- **ML Algorithm Training and Test Data Analysis:** Predominantly Reetikaa Gajula with help from Mark Jeakle and Sylvia Miller
- **Visualization, Result Analysis, Conclusion:** Equal Contribution
- **Final Tutorial:** Predominantly Mark Jeakle and Reetikaa Gajula with help from Sylvia Miller
- **Final Report:** Predominantly Sylvia Miller with help from Reetikaa Gajula and Mark Jeakle
- **Tutorial Video:** Equal Contribution

Special Note: All members of this group contributed equally and assisted each other wherever necessary. All members contributed equally to data exploration and summary statistics, visualization, result analysis, conclusion, and the final report, tutorial and video. Since author contributions were equal, the ordering of the authors is in alphabetical order by last name.

- **Reetikaa Gajula:** Reetikaa contributed her insights and experience with machine learning algorithms, and greatly assisted the group in the design, development, and training of the machine learning algorithm. Reetikaa assisted in the visualization, result analysis and conclusion of the project.In addition to this, she also contributed to creating a user friendly final tutorial and assisted in ensuring complete clarity for the final report.
- **Mark Jeakle:** Mark contributed his expertise and creativity by means of idea creation, dataset preprocessing and data exploration. Mark's work with his initial idea helped the project run smoothly and assisted the team in forming an initial hypthoesis. Mark assisted in the visualization, result analysis and conclusion of the project. In addition to this, he contributed greatly to creating a detailed final tutorial and ensuring clarity for the final report.
- **Sylvia Miller:** Sylvia contributed her experience with machine learning algorithms and data visualizations by assisting the group in the dataset preprocessing, exploration and machine learning testing and analysis. Sylvia assisted the group greatly by summarizing the findings of the project in the final report and assisting with ensuring data visuals were meaningful. Sylvia's experience publishing in IEEE assisted the group in a thoughtfully formatted final report.

Analyzing S&P 500 Performance and U.S. Presidential Election Outcomes

Reetikaa Gajula
Science Academy
University of Maryland
College Park, MD
rgajula@umd.edu

Mark Jeakle
Science Academy
University of Maryland
College Park, MD
mjeakle@umd.edu

Sylvia Miller
Science Academy
University of Maryland
College Park, MD
smille30@umd.edu

***Abstract*—During this politically turbulent time, this project explores the relationship between stock market performance by examining the S&P 500 index compared to the outcome of the US presidential election. Historical data from the S&P 500 index and election results were combined and examined to analyze potential correlations, differences in stock trends under different political parties (e.g., Democrat and Republican) and were trained to determine whether stock market performance could be used as a predictive power for election results. Early examination of data revealed a weak correlation between stock performance and election outcomes. Despite these early observations, the project continued to create a machine learning model to attempt to use the stock market as a predictor for the election outcome. Our model was ultimately unsuccessful, with all models predicting a Democratic win.**

***Index Terms*—S&P 500, Presidential Election, Weak Correlation**

I. Introduction

Being located in the Washington, DC area, it is difficult to avoid political discussions, especially during a presidential election year. Due to this, our team frequently heard varying hypotheses regarding the correlation of market performance trends and the outcome of the US presidential election. Our team internalized this hypothesis and sought to determine its validity through data science and even generating three different machine learning models to predict the outcome of the US presidential election.

Understanding that elections often impact financial markets, our team included historic S&P 500 data 1997 to 2024 as well as historic election results data from 1976 to 2020 focusing on the 1997 to 2024 time frame. Our team felt the S&P 500 index was an adequate approximation of stock market performance trends without the need for making our dataset incredibly large thus increasing our computational load. Included within the analyzed date range is COVID-19. This is helpful to visualize the impacts of volatile economic times and provides variance in the dataset.

Determining the correlation between stock market performance trends with US presidential elections is important since there is a strong prevalence of confirmation bias within the media and public. After the presidential election this year, there were many news articles and public expression that the market improved after the election of Republican nominee Donald Trump. While the stock market performance did increase, our team examined the data in order to determine if this was correlational with a Republican nominee or if the stock market improved at a rate that was not abnormal to the rater observed with the election of a Democratic nominee.

The null hypothesis our project tested was that the outcome of US presidential elections have no statistically significant impact on the financial markets as indicated through the trends portrayed in the S&P 500 index. While elections nationally as well as worldwide are often portrayed to impact financial markets, our team seeks to determine if this impact is statistically significant from the average trend of the market. Furthermore, our team seeks to leverage insights gained through performing data science of historical S&P 500 data and presidential election data to create three machine learning algorithms to be used in predicting the outcome of the US 2024 presidential election between Democratic nominee Ka-

mala Harris and Republican nominee Donald Trump based on the S&P 500.

The project aims to:

- Explore historical S&P 500 trends and their relationship to presidential election outcomes.
- Investigate differences in stock performance based on the winning political party.
- Determine the feasibility of using stock data as a predictive tool for elections.

II. Methodology

A. Data Collection

The project analyzes S&P 500 historical stock price data obtained in HTML format from Yahoo Finance from 1997 to 2024. The dataset contains daily metrics such as date, open, close, high, low, and volume. The dataset is large enough to span several years capturing financial market trends over variance turbulent times and provides helpful insight to average performance metrics.

The US presidential election data from 1976 to 2020 sourced from Harvard, which contains metrics for year, candidate, party, votes, and state. The election data is fairly straightforward and contains data every 4 years since the US presidential election is conducted on a 4 year cycle.

Both datasets were selected for their clarity, near-completeness and relevance to our hypothesis testing. The data was stored in a shared Google Drive for group to work on together and then processed using a shared Google Colab.

B. Data Preparation/Preproccesing

In order to begin data preparation and preprocessing the team first conducted HTML parsing of the S&P 500 using `BeautifulSoup`, and focused on the key metrics of date, open, high, low, close, adjusted close, and volume. The data was then extracted into a structured csv format and saved as a data frame for ease of SQL use to query.

Although the S&P 500 dataset was already fairly cleaned, the team ensured thoroughness by converting the date column to the datetime format to facilitate the time series analysis. Additionally, non-numeric characters such as commas from columns like volume and close were removed and converted to numeric types. Lastly, to handle missing values in the dataset, the values were forward-filled (`ffill`) to preserve continuity in time series data. Any rows where the date values were missing, and any duplicates were dropped from the dataset.

After ensuring that the dataset was cleaned and could be used in conjunction with the US presidential election dataset, the two sets of data were merged based on aligning the year ensuring that only matching records were retained. Lastly, after a quick examination of the data, the team feature engineered annual percent change in the S&P 500 closing prices, and party vote share in the US presidential election data in order to compare Democratic and Republican dominance.

Primary Data Preparation Steps:

- HTML Parsing: S&P 500 data was extracted using `BeautifulSoup`, focusing on key metrics: Date, Open, High, Low, Close, Adjusted Close, and Volume.
- Data Cleaning:
  - Missing and invalid entries were handled through forward-filling.
  - Non-numeric characters (commas) were removed from financial columns (Volume, and Close).
  - Duplicates were dropped.
- Integration: The cleaned S&P 500 data was merged with US presidential election data by matching the year field.
- Feature Engineering:
  - Annual Percentage Change: Calculated from S&P 500 closing prices
  - Party Vote Share: Created from the US Presidential Election dataset

C. Exploratory Data Analysis

The most apparent observation from the exploratory data analysis was that there is a steady growth in the S&P 500 index over the decades with obvious dips during major economic events such as the 2008 financial crisis and COVID-19.

When performing correlation analysis, we identified a weak positive correlation of 0.0246 observed between S&P 500 closing prices and the total votes received by candidates. After further exploratory analysis we created heatmaps which revealed very minor correlations between financial metrics and vote shares.

Ultimately, our exploratory analysis suggested that our conclusion of this project would be to fail to reject our null hypothesis.

Summary of Conducted Exploratory Analysis:

- Correlation Analysis: Examined the relationship between S&P 500 closing prices and candidate votes using scatter plots and Pearson correlation.
- Party-Based Comparisons: Compared average S&P 500 closing prices and trading volumes under Democratic and Republican administrations.
- Hypothesis Testing: Used t-tests to determine the significance of differences in trading volumes between political parties.

D. Visualization

Exploring the relationship between S&P 500 closing prices and candidate votes reveals a very minimal correlation between S&P 500 Closing Price and Candidate Votes: 0.0246. As displayed in Figure 1 below, the market consistently trends upwards and visually has no significant impact on candidate votes.

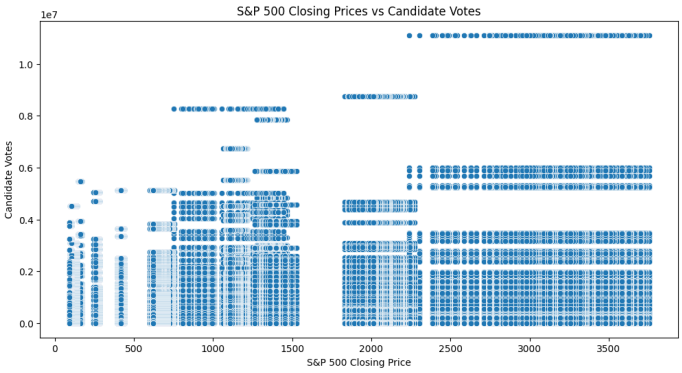

Fig. 1. S&P Closing Prices vs. Candidate Votes

After conducting a correlational analysis that revealed a marginal correlation between the closing price of the S&P 500 and candidate votes. We examined the difference in closing prices after the election of a Democratic candidate compared to the election of a Republican candidate. The plot of the results did not show a significant delta between the average closing price in either political party. The average S&P 500 closing price under Democratic wins is $1019.41 compared to the average S&P 500 closing price under Republican wins is $1017.48. As shown in the bar graph below, the delta between closing prices for each political party is negligible.

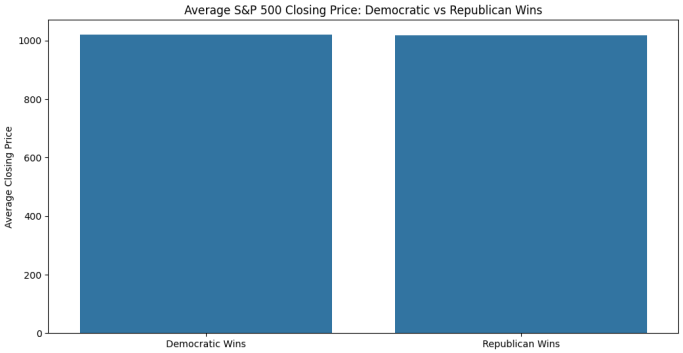

Fig. 2. Democrat wins: $1019.41 vs. Republican wins: $1017.48

Examining the distribution of the S&P 500 volume for Democratic vs. Republican wins reveals that the Democratic and Republican values are overlayed as symbolized by the purplish hue of Figure 3 below. Performing a T-Test for Volume under Democratic vs Republican Wins produces the T-statistic of 0.40827, and a P-value of 0.6830.

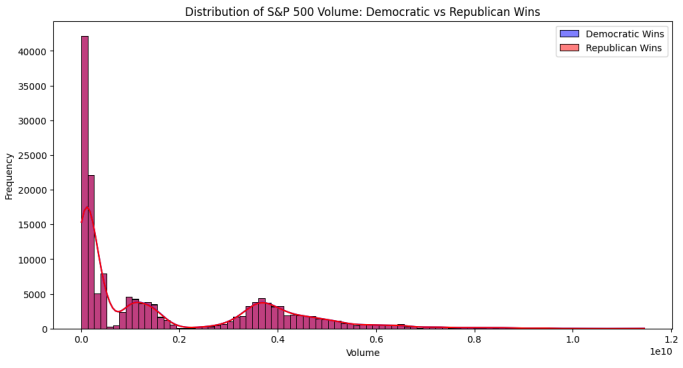

Fig. 3. Distribution of S&P 500 Volume: Democratic vs. Republican wins

Next, our group plotted the S&P 500 performance vs party vote share below in figure 4, revealing a very even split across Democratic and Republican vote shares as it

is related to the S&P 500 performance. When calculated, the correlation between S&P 500 change and Democrat vote share is 0.23, and the correlation between S&P 500 change and Republican vote share is -0.23.

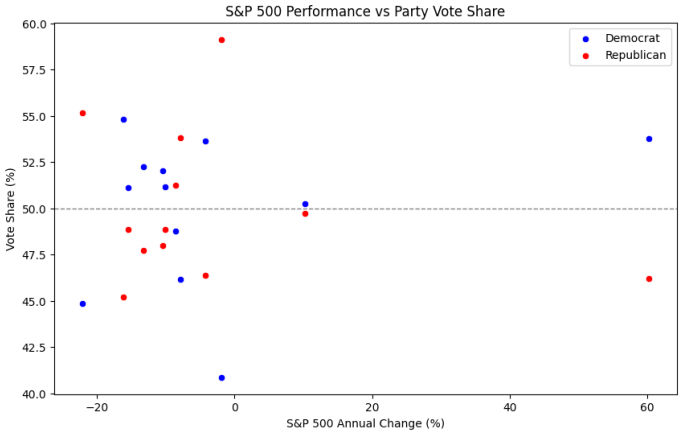

Fig. 4. S&P 500 Performance vs. Party Vote Share

Moving on, our team plotted in Figure 5, the linear regression of the S&P 500 annual change vs. party vote share for understanding correlation and trends. The calculated Democrat regression coefficient is 0.04426 and the calculated Republican regression coefficient is -0.04426.

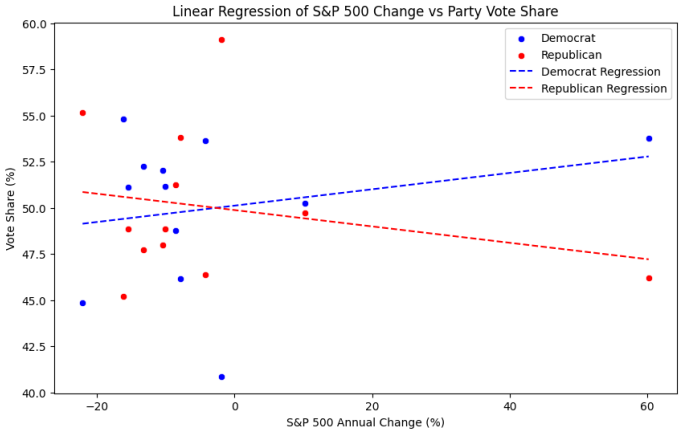

Fig. 5. Linear Regression of S&P 500 Change vs. Party Vote Share

We visualized trends in vote shares and S&P 500 annual changes over time using a line chart to show the trend in Figure 6. As can be observed from the line chart, there appears to be only one peak in S&P 500 change, and visually, the political parties share dominance switches off without any obvious alignment to the S&P 500.

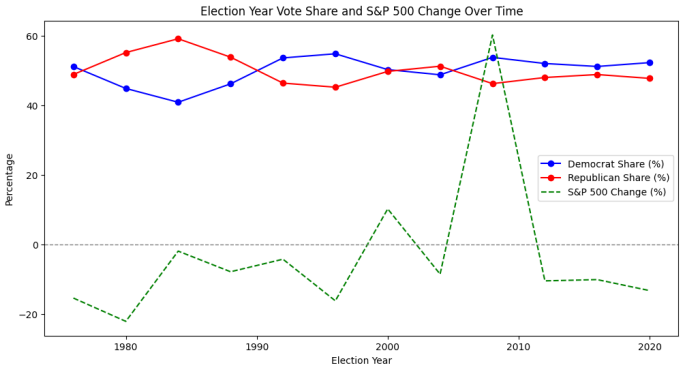

Fig. 6. Election Year Vote Share and S&P 500 Change Over Time

Lastly, visualizing the Logistic Regression Decision Boundary shown in Figure 7 below reveals where the decision boundary lies for the S&P 500 annual change for the determination of the winning party.

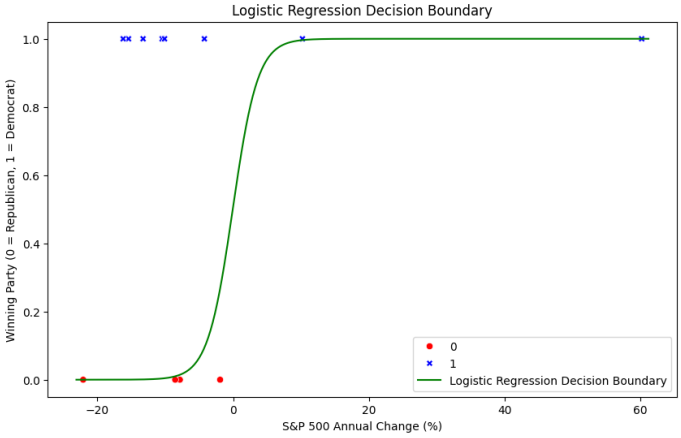

Fig. 7. Logistic Regression Decision Boundary

E. Machine Learning Analysis

In order to perform machine learning on the datasets in order to predict the outcome of the 2024 US presidential election based on S&P 500 trends, we imported 3 primary machine learning libraries: Scikit-learn, TensorFlow, and XGBoost (short for eXtreme Gradient Boosting).

The Models Explored are:

- Logistic Regression
- Decision Tree

- Random Forest
- Gradient Boosting with XGBoost

One of the machine learning methods we explored was **Logistic Regression**, which is typically employed as a baseline model for binary classification tasks. This works well for our task as our goal is to create a prediction of either Republican or Democrat, where the logistic regression method ultimately provides an estimation of a categorical dependent variable. The advantages to this model are its simplicity and interpretability as well its robustness against overfitting for low dimensional data. The primary limitation to this method is the creation of linear decision boundaries where the data may not be linearly separated and therefore may not capture the necessary patterns for successful classification.

Next, we used the **Decision Tree**, which is a classic fundamental model that provides ease of interpretability and limited sucess for classification tasks. While our team did not have high expectations for this model, we were curious as to the evaluation metrics for something that we could easily comprehend and interpret afterwards. This model builds a tree structure by splitting the data into subsets based on features aiming to increase information gain and decrease entropy. Each node in the tree represents a decision rule and each 'leaf' represents an outcome. We selected a `max_depth` to be 3 in order to prevent overfitting. The advantages of this model are its ease in interpretation, its capability to work with both categorical and numerical data, and its ability to work with non-linear relationships as opposed to the previously discussed Logistic Regression method. The limitations are that the model is prone to overfitting, which is why we specified the maximum depth of the tree, and the performance is sensitive to outliers.

Building off of the Decision Tree, we then examined **Random Forest**, which is an ensemble method implemented for its ability to handle non-linear relationships between features. The ensemble aspect of Random Forest involves the aggregation of predictions for multiple decision trees in order to improve accuracy and reduce overfitting. For our model, we utilized 100 decision trees determined by the `n_estimators` hyperparameter. The advantages of this method are the typical improved accuracy compared to the standard decision tree due to the ensemble averaging, and the model's robustness to outliers and overfitting. The limitations of this model are that it is computationally intensive, and there is reduced interpretability especially as compared to the decision tree model.

Lastly, our team explored **Gradient Boosting** with XGBoost. The choice of using XGBoost was made for its asvanced boosting techniques, which iteratively improve the model by making improvements on the errors made by previous iterations. This model was convenient to build and use due to the built-in regularlization parameters that prevent overfitting. The advantages of this method is the model's ability to handle missing data, and its typical improved performance. The limitations are that it can require a lot of careful tuning of the hyperparameters for the optimal performance outcome our team seeks, which is difficult to do quickly and with ease. Additionally, this model can overfit on small datasets if there are not proper regularizers.

III. Insights & Conclusion

In conducting the correlation analysis between the S&P 500 trends and votes, our team found minimal to no correlation. Additionally, our team found no significant difference between the winning political party and the S&P closing price. Since there was minimal to no correlation, our team found it unlikely to fail to reject our null hypothesis which was that the outcome of US presidential elections have no statistically significant impact on the financial markets as indicated through the trends portrayed in the S&P 500 index.

Despite initial observations through exploratory analysis revealing an unlikeliness for the two datasets to be correlated. Our team built and trained four machine learning models: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. This was still conducted to determine if there were any features within the dataset that were contributing to a correlational relationship that was not observed through our exploratory analysis.

Ultimately, our model was unsuccessful in correctly predicting the outcome of the 2024 US presidential election. Every model predicted that the outcome would be the Democratic candidate, who in the case of this election is Kamala Harris. However, the Republican candidate, Donald Trump, won the election. Additionally, the model s predicted the Democrat vote share to be

50.34% and predicted the Republican vote share to be 49.66%.

Final Results:

- Logistic Regression: Democrat
- Decision Tree: Democrat
- Random Forest: Democrat
- XGBoost: Democrat
- Predicted Democrat Vote Share: 50.34%
- Predicted Republican Vote Share: 49.66%

From our models' results, our team has concluded that we fail to reject our null hypothesis. In conducting this project our team has proved that there is no correlation between the stock market's performance as portrayed by the S&P 500 on the US presidential election. Interestingly, while our team conducted this analysis there were many headlines such as "Stocks soar after Trump defeats Harris in election 2024 outcome" from Fox Business, and "Why Donald Trump's election win fuelled a stock market surge" from Queen Mary University of London. Our project has proven that these articles hold no validity, and that while the stock market may have improved this has no correlation to the election of a specific candidate or political party.

IV. Data Science Ethics

Considering how politically charged the United States has become and our proximity to the Washington, DC area, our team had several ethical considerations for this project. Our team ensured that there was no inherent biases reflected in the dataset by sourcing each dataset from a refutable source. Additionally, while we cannot portray every class, our team focused on portraying an average representation of the financial market using the S&P 500 data which impacts the entire country at large. This could have been a problem if our dataset only included the financial trends from specific regions or demographics not allowing for generalization for the entire population.

Our next consideration was having transparency in feature selection. Our team visualized the data extensively showcasing the different ways to interpret that data in order to show if there is any possibility for correlation between the datasets.

Lastly, our team was careful to present the data in a way that is not political despite the project's relation to politics. This straightforward presentation of data was intentional as to not isolate a certain group or region.

Ultimately, our team did our best in ensuring that ethical considerations were made in conducting this project.

References

[1] MIT Election Data and Science Lab. (2017). *U.S. President 1976–2020* [Data set, Version 8.0]. Harvard Dataverse. https://doi.org/10.7910/DVN/42MVDX

[2] Yahoo Finance. (2024). *S&P 500 Index Historical Data*. Retrieved from https://finance.yahoo.com/