Project Report

Modeling US. Presidential Election Using Regression Analysis and Machine Learning Algorithms

By:

Dhiaa Eddine Grar

June 2024

Abstract

The aim of this project was to conduct a thorough study of US Presidential Election Modeling techniques. Initially, previous literature on this topic was reviewed both theoretically and empirically to gain insight into the methods and findings of previously developed models for predicting election outcomes. Subsequently, this study developed new models based on the existing methodologies to understand the relationship between election outcomes and various economic and non-economic factors. These models were evaluated and tested to assess their performance. Ultimately, the developed models were used to predict future outcomes, such as the results of the 2024 presidential election. Examining existing literature was a crucial step in building new models, as it provided valuable references for this project. This project used both regression and classification techniques to build forecasting models. Data was collected manually from official websites of the American government and various American organizations. Election outcomes were used as the dependent variable, and several variables inspired by previous research were considered in this study. The research models were constructed and evaluated. Finally, this study found that both regression and classification techniques could be effective options for modeling US presidential election results.

Executive Summary

This study explores the relationship between US presidential election outcomes and various economic and political factors. To understand the connection between voting behavior, economic indices, historical patterns, and other political measures, this study reviews a variety of previously constructed forecasting models on this topic. Then uses a sample of election cycles ranging from 1904 to 2020 to build new models based on variables identified by economists, researchers, and political scientists.

By considering the voting results of the incumbent party as the dependent variable, and independent variables that outline economic performance and other political and geopolitical indicators, the objective is to comprehensively assess the relationship between these variables. The research methodology employs both regression and classification techniques to study this topic, and the proposed hypotheses are evaluated using appropriate evaluation metrics for each model.

The findings reveal that while regression models, even with some error margin, can provide valuable insights into the predominance of a candidate, classification models offer high accuracy but lack a precise view of the candidate's advantage. The study suggests that the discussed variables can explain election outcomes effectively, indicating that a forecast of future results would be valuable and reliable.

Introduction

The United States presidential election is one of the major events that affect the local and worldwide economies. Along with their impact on the global economy, US elections also have an impact on the growth of competitive local markets. The importance of this event lies in its potential to impact American domestic policies as well as other countries' policies, agreements made with foreign nations, and political dynamics on a global scale. A new U.S. president can drive significant changes that have repercussions on international trade, economic partnerships, and diplomatic relations. These shifts can affect everything from global market stability and environmental policies to security alliances and international cooperation on critical issues such as climate change and public health. Consequently, the outcome of the U.S. presidential election holds substantial weight in shaping the future direction of both the United States and the world at large scale. Given its significance, it is crucial for researchers to develop accurate predictive models for election outcomes. Such forecasts can provide valuable insights for countries and institutions, enabling them to adjust policies in response to potential shifts in U.S. presidency. The importance of the result of the United States presidential election is well known among the major developed and developing economies worldwide. Election forecasts provide governments, businesses and international organizations with early insight about change in the U.S policies directing them to make suitable change in their policies. They can eliminate negative impacts in the economy, strengthen international relations, and contribute to political stability at the international level. In anticipation of the outcome of the election, various political scientists and economists worldwide have been trying their hands at predicting the election result.

In most cases, modeling an election is complex. Because it includes various factors and several candidates. Also the existence of many parties adds to the difficulty. However, the U.S. presidential election is rather easier to model due to its bipartisan system. Only two parties prevail the political field, the Republican and the Democratic. In this system, the success of one party typically means the failure of the other. Therefore, many forecasts use the incumbent party's votes as the dependent variable. This approach is based on the theory that the U.S. presidential election are highly affected by the incumbent party's performance during its tenure. Satisfied voters tend to support the incumbent party's candidate, while dissatisfied voters prefer the opposing party's candidate. A number of experts have developed models on this matter, starting from Ray C. fair (1978) who pioneered forecasting US presidential elections, followed by Lewis-Beck and Rice (1982), Alan Abramowitz (1984), Campbell and Wink (1990), Wlezien and Erikson (1996) Hibbs (2000) and Cuzán & Bundrick (2005) among many others. These political scientists continuously refined their models over time by introducing or removing predictive variables to enhance their performance.

However, most of the previous empirical work have omitted some important aspects when studying the relationship between the voter's behavior and various factors. Additionally, the majority of these studies were implemented on a limited dataset. Therefore, it is of interest to study this topic using all possible features able to explain the nature of the results, with the largest possible scale of data.

This report offers a thorough review of the existing literature on modeling U.S. presidential elections, highlighting their methodologies and findings. Building on these foundations, new forecasting models will be proposed which contain a wider range of variables and use a larger data scale. Data will be collected from various reliable sources to identify key predictive variables for constructing the models. These models will go through evaluation and testing to ensure their accuracy. The ultimate goal is to predict future election outcomes, starting with the 2024 election cycle. This approach aims to give reliable and precise forecasts.

The first chapter of this project will conduct a literature review by initially defining the concept of the US presidential election, followed by a theoretical and empirical review of previous works. The second chapter will present the methodology used in this research and define each model employed. Subsequently, the third chapter will focus on the implementation of the project processes, presenting the results and evaluating the performance of the models. Additionally, predictions for the 2024 election will be provided using each developed model. Finally, the project will give a conclusion that summarizes the findings, along with limitations faced during this research.

1. Literature review

Previous work on this topic will be reviewed and introduced. This section of the report will focuse on the numerous theoretical works forecasting the U. S. presidential election throughout history. A variety of methods and theories was employed to accurately predict the winner of the election. An overview of these studies involving their methodologies will be listed. Thus, this study will form the basis of building new models for forecasting the next elections.

1.1 Concept of US presidential elections

The U.S. presidential election, held every four years, is a pivotal event that significantly impacts the country's domestic and global political and economic landscapes. For several decades, two major parties, the Democratic and the Republican, have dominated these elections. Each party selects its candidate through a series of internal nomination processes. And ultimately, two major nominees emerge to compete for the presidency.

During the general election, the electorate casts their votes for the candidate they believe is most suitable. However, the final decision is made by the Electoral College, a body of representatives from each state who vote on behalf of their constituents. The number of Electoral College representatives for each state is equal to the sum of its Senators and House Representatives, as shown in figure (1.1), reflecting the state's population size.

While voters in each state cast their votes for their preferred candidate, they are actually voting for group of electors who have pledged to support that candidate. Generally, the candidate who wins the popular vote in a state receives all of that state's Electoral College votes, following a "winner-takes-all" system.

This alignment typically means that the Electoral College vote corresponds with the popular vote within each state, thus reflecting the choice of the state's citizens.

The candidate who receives a majority of the Electoral College votes (at least 270 out of 538) wins the presidency. This system emphasizes the importance of winning key states with significant electoral votes and explains why candidates often focus their campaigns on swing states. The U.S. presidential election process, combining popular and electoral votes, ensures that both the will of the general population and the federal structure of the nation are considered in the final outcome.

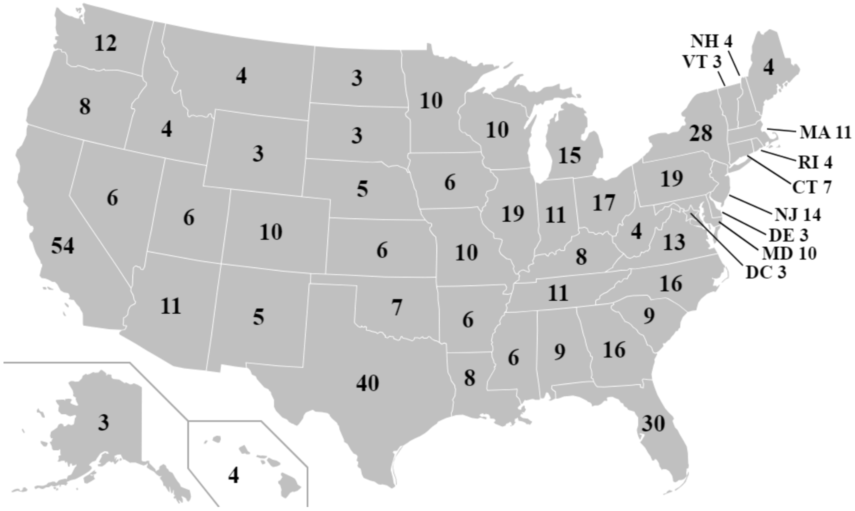


Figure 1.1: US Electoral College votes per state

1.2 Theoretical Review

1.2.1 Theoretical review of OLS modeling techniques

Over the last few decades, economists and political scientists have developed a variety of regression models to forecast US presidential elections. A number of these models have been abandoned, a few have seen significant changes, and a few have seen very little change since their creation. Tracing this historic process will demonstrate how later attempts have built on earlier research and evaluate the accuracy of the regression models which are currently in use, whose structural elements have remained mostly unchanged. Although forecasting has been widely used in many other fields, it is still relatively recent in the area of political science. Here are some prominent economists and the models they have proposed:

- ***Ray C. Fair (1978):*** An economist who developed the first econometric model that forecasts US presidential elections based on economic factors like GDP growth and inflation. He attempted to include factors such as the economic condition of the election year and the incumbent party in the context of the forecasting model. He believes that people will vote for the incumbent party's presidential candidate or another one if the economy is doing well. In the case that the economy fails to perform well, people will probably elect the presidential candidate of the opposing party. Most of the current models for predicting presidential elections have derived from his study. In 1996 he refined the model. The variables included were GNP, inflation rate, and the consecutive terms in which the incumbent party governs the country. The dependent variable was the percentage of the incumbent party votes. Another indicator that highlights wartime periods was later added by Fair to his revised model
  Publication which contains the model: *"Predicting Presidential Elections and Other Things"*

- ***Lewis-Beck and Rice (1984):*** Political scientists Michael S. Lewis-Beck and Tom W. Rice introduced a model that is considered a seminal contribution to the field of political forecasting, particularly in predicting U.S. presidential elections. This model incorporate various economic indicators, such as inflation rates and income growth, with incumbency status. Their contribution consists of using the president’s job approval rating as an independent variable. This rating is obtained through a poll conducted on a sample of people from the population, where they are asked whether they approve or disapprove of the president in office. Lewis-Beck and Rice expanded their first model by including two other variables: the strength of the incumbent party's presidential nominee in the primaries (elections to determine the final candidate of each party) and the party’s performance in previous midterm elections.
  Publication which contains the model: *"The American Voter Revisited"*

- **Alan Abramowitz (1988):** A political scientist known for developing the "Time for Change" model, which forecasts election results using economic indicators and the president's approval rating. Abramowitz added a "time for a change" variable into the Lewis-Beck and Rice model, which represented voters' cyclical tendency to switch parties in the White House every eight years. He

used the ordinary least-squares (OLS) method to estimate the coefficients of the linear regression model.
Publication which contains the model: *" The Time-for-Change Model and the 1992 Election"*

- ***Campbell and Wink (1990):*** Political scientists James E. Campbell and Kenneth A. Wink introduced a model that incorporates both political and economic variables. Campbell and Wink adopted the earlier strategy proposed by Lewis-Beck, 1984, by pairing campaign polls with economic growth.
Publication which contains the model: *" The American Campaign: U.S. Presidential Campaigns and the National Vote"*

- ***Wlezien and Erikson (1996):*** Political scientists who presented a model with the percentage of votes for the incumbent party and economic indices as variables. To assess the precision of their forecasting model, they employed the adjusted R square and the R square (coefficient of multiple determination). Subsequently, Erikson and Wlezien (2016) improved their model by incorporating the polls into their equation.
Publication which contains the model: *"Temporal Horizons and Presidential Election Forecasts,"*

- ***Hibbs (2000):*** Douglas Hibbs, an economist, presented his model to forecast the results of U. S. presidential elections in 2000. Hibbs's model is also known as the "Bread and Peace" model because it emphasizes the impact of economic performance and military conflict on electoral outcomes. Particularly, it hypothesizes that the incumbent party's vote share is affected by real disposable personal income growth per capita (the "bread") and the cumulative number of American military fatalities in overseas conflicts (the "peace"). This model has been used to show how warfare and economic conditions in particular have a big influence on presidential elections.
Publication which contains the model: *" Bread and Peace Voting in U.S. Presidential Elections,"*

- ***Cuzán & Bundrick (2005):*** Economists Alfred G. Cuzán and Charles M. Bundrick proposed a model for forecasting U.S. presidential elections in 2005 known as the "Fiscal Model" which focuses on the impact of fiscal policy on election outcomes. It suggest that the incumbent party's vote share is influenced by changes in real federal expenditures as a proportion of GDP. Their theory suppose that voters respond negatively to excessive government spending. This model highlights how fiscal policy affects voter behavior and election results.
Publication which contains the model: *" Fiscal Policy and Presidential Elections: Update and Extension,"*

The following table (1.1) shows the independent variables used by each model from the discussed OLS models:

	Fair (1978)	**Lewis-Beck and Rice (1984)**	**Alan Abramowitz (1988)**	**Campbell and Wink (1990)**	**Wlezien and Erikson (1996)**	**Hibbs (2000)**	**Cuzán & Bundrick (2005)**
Economic Growth	**X**	**X**	**X**	**X**	**X**	**X**	**X**
Inflation	**X**	**X**					**X**
Incumbent Party	**X**					**X**	**X**
Incumbent President	**X**						
Presidential Approval		**X**	**X**		**X**		
Candidate Preference (polls/surveys)				**X**	**X**		
War Occurring/ Military Deaths	**X**					**X**	**X**
Federal Spending							**X**
Terms in office	**X**					**X**	**X**
2-term Penalty Cycles			**X**				
Incumbent Party Results, House Elections		**X**					
Incumbent Share of Primary Vote		**X**					

Table 1.1: Historical OLS models and their associated predictors

1.2.2 Theoretical review of other modeling techniques

- ***Logistic regression model of Nate Silver***: Nate Silver is a statistician and the founder of FiveThirtyEight, a website specializing in data and statistical analysis, especially in the field of political forecasting. Silver is well known for his innovative use of statistical models to predict U.S. presidential elections. Logistic regression is one of the foundations of his approach, a technique he applies to estimate the probability of a candidate winning based on a range of variables. In his model, the dependent variable is the binary election outcome of each candidate (win or lose). While the independent predictors include national and state polling data, economic indicators such as GDP growth and unemployment rates, presidential approval ratings, demographic information, and historical voting patterns. Despite the complexity of Nate use of logistic regression, his models have continuously generated accurate forecasts, Because it incorporate several important factors, adapts dynamically to new data, and provides predictions that take into account sources of uncertainty. Silver's models have proved to be reliable in offering accurate prediction of election results thus enhancing the credibility and reliability of election forecasting.
  Publication which contains the model: *" The Signal and the Noise: Why So Many Predictions Fail—but Some Don't"*

- ***Lasso regression model of Pankaj Sinha and others:*** A research paper published in 2020 that predicts the US presidential election results using Lasso regression. Pankaj Sinha and other researchers collected data and divided variables into two categories. A category of economic variables that includes inflation, unemployment rate, exchange rate, and economic growth rate. And another category of Non-economic variables which are president job approval rating, crime rate, number of terms, campaign spending index, midterm performance, and scandal rating. They developed a Lasso regression model to enhance their OLS regression model published in earlier research paper using the same variables.
  Publication which contains the model: *" Prediction for the 2020 United States Presidential Election using Machine Learning Algorithm: Lasso Regression"*

- ***Artificial Neural Networks Model (ANN) model of Mohammad Zolghadr:*** A research paper published in 2018 that aims to develop an accurate model for predicting the US presidential election with Artificial Neural Networks algorithm. Researcher Mohammad Zolghadr used the electoral votes of the incumbent party as the dependent variable as it will determine who will win the election. The data used ranged from 1952 to 2012. And independent variables considered are: Personal income, Electoral votes of the incumbent party in the previous election, Votes of the incumbent party in the last senate election, Votes of the incumbent party in the last house of representative election, the president's approval rate, Unemployment rate and the monthly GDP.
  Publication which contains the model: *" Modeling and forecasting US presidential election using learning algorithms"*

- ***Bayesian Model of Brittany Alexander:*** A scientific article published in 2017 by researcher Brittany Alexander. It developed a Bayesian model in order to predict the winner of the US presidential election in each State. In this model, the final percentage of votes that each candidate receives in a specific state is the dependent variable. While the poll results, their mean and variance and the polls timings are the independent variables. The model uses a Bayesian approach by considering poll data from states that are similar as priors. The accuracy of the election outcome forecasts is improved by the combination of contextual data from similar states with direct polling data.
  Publication which contains the model: " *A Bayesian Model for the Prediction of United States Presidential Elections"*

- ***Decision Tree model of Joseph Stoffa and other researchers:*** This model was published in a scientific article in 2018. It used a decision tree algorithm to predict county-level voting based on historical patterns, specifically the percentage of elections a county voted Republican. The decision tree, using either Gini impurity or information gain, used several independent variables: historical voting records, the number of past elections considered, socio-economic data (such as median age, ethnic makeup, crime rates, and unemployment rates), and party platforms. Data was gathered from the 2004, 2008, and 2012 elections. Initially, the model was trained only on voting records and past elections. Later, socio-economic data and party platforms were added to improve prediction accuracy. The final model was evaluated on its ability to predict the 2016 election. This approach showed that including socio-economic factors and party platforms significantly enhanced the model's accuracy in predicting election outcomes.
  Publication which contains the model: " *Predicting How U.S. Counties will Vote in Presidential Elections Through Analysis of Socio-Economic Factors, Voting Heuristics, and Party Platforms"*

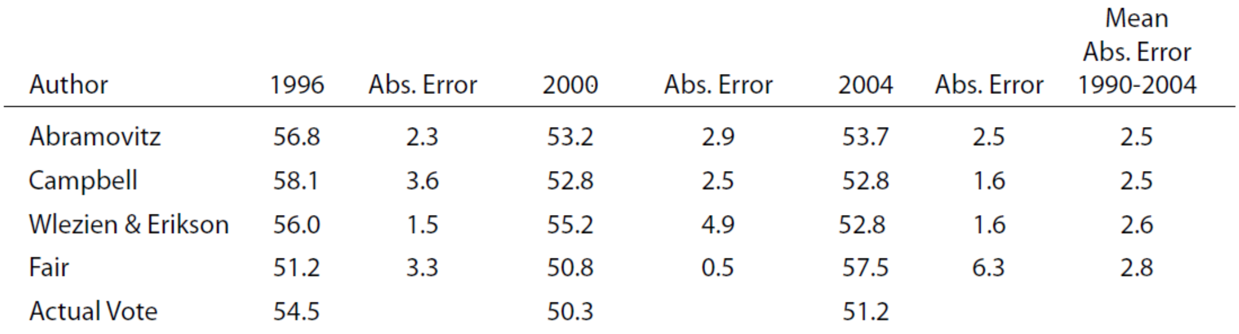
1.3 Empirical Review

Throughout history, several political scientists, economists, and researchers have attempted to build models for predicting U.S. presidential elections. Each new study builds on the previous ones, introducing innovative approaches and incorporating new variables to enhance the accuracy of the forecasts, with taking advantage of the latest available data. While early models primarily utilized Ordinary Least Squares (OLS) regression techniques, the use of advanced machine learning algorithms has increased in recent study. These modern researches intend to offer greater predictive power and flexibility and contribute significantly to the field of election forecasting by capturing complex patterns and relationships that traditional methods might miss. This section aims to look at the findings of the models discussed earlier and evaluate their performance and accuracy.

1.3.1 Empirical results of OLS modeling techniques

OLS regression models have demonstrated satisfactory results with a small error margin. One significant advantage in this type of modeling is that each new model builds on its predecessors, benefiting from their hypothesis and conclusions. This iterative improvement process allows researchers to refine their methods continuously. The objective is that the contributions provided by new models will yield even more accurate estimates of election outcomes.

The first model developed in this matter was by Ray Fair (1978). This model emphasized on a small set of economic variables that includes GDP growth and inflation. Also it incorporates the number of president's terms and the incumbency status. Despite being the first attempt in the field, Fair's model has shown a reasonable degree of accuracy and a provided valuable insight into the impact of economic conditions on presidential election results. With this model, Fair predicted correctly the winner of 1980 election, and anticipated Ronald Reagan's victory over incumbent Jimmy Carter. He also correctly predicted the winners for 1984 and 1988 election cycles. And in his fourth prediction, Fair's model failed to identify the upcoming president, incorrectly predicting the reelection of President George Bush. This failure led Fair to revise his model. Subsequently other political scientists and economists proposed new models that have derived from Fair's fundamental model, such as Abramovitz, Wlezien & Erikson, and Campbell. The following table (1.2) shows the prediction results of each of these models for the three election cycles: 1996, 2000, and 2004.



Author	1996	Abs. Error	2000	Abs. Error	2004	Abs. Error	Mean Abs. Error 1990-2004
Abramovitz	56.8	2.3	53.2	2.9	53.7	2.5	2.5
Campbell	58.1	3.6	52.8	2.5	52.8	1.6	2.5
Wlezien & Erikson	56.0	1.5	55.2	4.9	52.8	1.6	2.6
Fair	51.2	3.3	50.8	0.5	57.5	6.3	2.8
Actual Vote	54.5		50.3		51.2		

Table 1.2: Comparing Forecasts of the Incumbent's Percent Share of the Two-Party Vote in Presidential Elections, 1996-200

These OLS models provided relatively accurate predictions with small error margins. Notably, the mean absolute error for the three election cycles is slightly lower in the other models compared to Fair's one. This indicates that these models have successfully enhanced the accuracy and reliability of the forecasts.

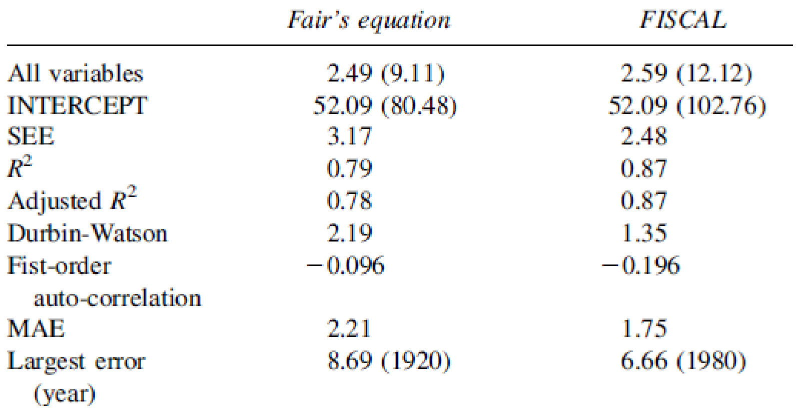
A number of Models have emerged based on the Fair's method by making various adjustments and refinements. Lewis-Beck and Rice emphasized survey data to capture voter intentions more precisely. Alan Abramowitz introduced a variable to take into consideration the number of terms a party has held the presidency. Campbell and Wink focused on the early economic indicators and their timing. Wlezien and Erikson integrated dynamic polling data into their model. Hibbs Used real disposable income growth and military casualties as predictors. Finally, Cuzán and Bundrick introduced a fiscal model that included government spending and budget balance. The following table (1.3) shows the prediction of 2008 election results by each one of these models compared with the actual result

Forecasting Model	**Forecast**	**Error** **(- is underforecast)**
Actual outcome was 53.4%.		
Ray Fair	51.5	-1.9
Michael Lewis-Beck	50.1	-3.3
Alan Abramowitz	54.3	0.9
Christopher Wlezien & Robert Erikson	52.2	-1.2
James Campbell	51.1	-2.3
Douglas Hibbs	53.7	0.3
Alfred G. Cuzán & Charles M. Bundrick	52	-1.4

Table 1.3: Regression Forecasts of the Democratic Party Share of the Two-Party Vote in the 2008 Presidential Election

All these models have demonstrated their reliability by accurately predicting the victory of the Democratic candidate Barack Obama in the 2008 presidential election. Each model forecasted a vote percentage exceeding 50% for the Democratic Party.

Alfred G. Cuzán & Charles M. Bundrick (2005) adopted the same Fair's Equation for their model known as the "Fiscal Model", with incorporating a new predictive variable named FISCAL. This variable represents the government's fiscal policy situation which is measured by the balance of the federal budget and indicates whether there is a surplus or deficit. The following figure (1.2) shows the evaluation metrics for their regression model compared to the Fair basic equation. The findings suggest that Alfred G. Cuzán & Charles M. Bundrick have slightly enhanced the performance of Fair's Model



	Fair's equation	*FISCAL*
All variables	2.49 (9.11)	2.59 (12.12)
INTERCEPT	52.09 (80.48)	52.09 (102.76)
SEE	3.17	2.48
R^2	0.79	0.87
Adjusted R^2	0.78	0.87
Durbin-Watson	2.19	1.35
Fist-order auto-correlation	−0.096	−0.196
MAE	2.21	1.75
Largest error (year)	8.69 (1920)	6.66 (1980)

Figure 1.2: Comparison of models: Fair's Equation vs. Cuzán & Bundrick Model

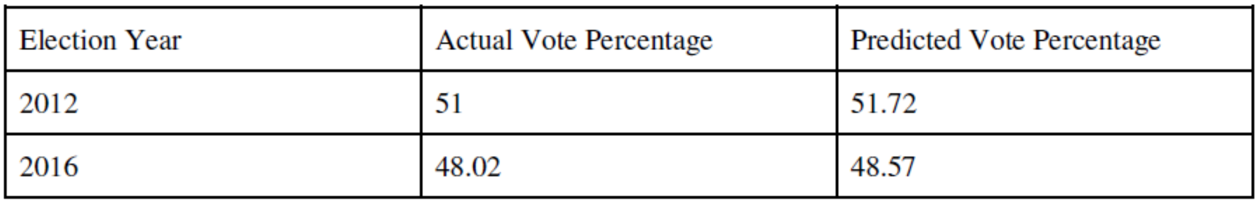
1.3.2 Empirical results of other modeling techniques

Logistic regression model of Nate Silver:

Nate Silver's use of the logistic regression method is highly sophisticated. He uses simulations to predict the election outcome millions of times. In 2012, Silver ran his model 10 million times. President Obama came out as the winner in 7.3 million simulations, while Mitt Romney won in 2.7 million times. Thus, Silver concluded that President Obama had a 73 percent chance of winning, as he won in 73 percent of the simulations.

Lasso regression model of Pankaj Sinha and other researchers:

The purpose of using lasso regression by these researcher was to refine their OLS model they developed in earlier publication. After testing the new model on the 2012 and 2016, the results obtained as shown in table (1.4) were highly accurate and it seems to capture the factors affecting the incumbent vote percentage.



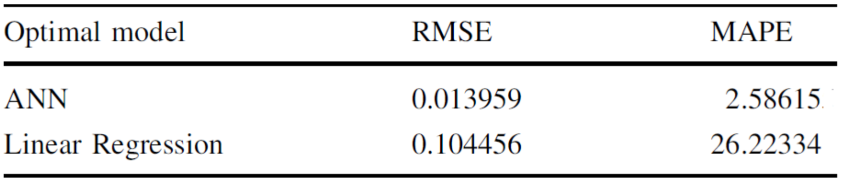
Election Year	Actual Vote Percentage	Predicted Vote Percentage
2012	51	51.72
2016	48.02	48.57

Table 1.4: Performance of Proposed Lasso regression model by Sinha for 2012 and 2016 elections

The proposed model predicted that the vote percentage for the incumbent party in the 2020 U.S. Presidential Election would be 41.63%. The actual outcome was 46.9%. Thus, the model correctly predicted the loss of the Republican candidate, Donald Trump, to Joe Biden.

Artificial Neural Networks Model (ANN) model of Mohammad Zolghadr:

Mohammad Zolghadr used the ANN algorithm approach to construct his model and compared it to a linear regression model he had previously developed. He founded that the ANN model exhibited significantly smaller RMSE and MAPE values compared to the linear regression model as presented in table (1.5), thus proving it to be a better model than linear regression.

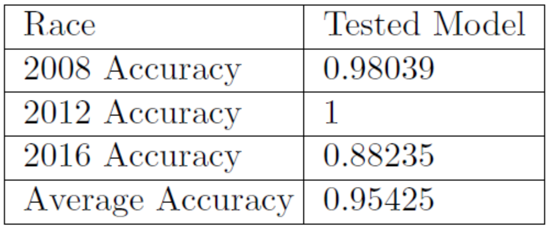


Optimal model	RMSE	MAPE
ANN	0.013959	2.58615
Linear Regression	0.104456	26.22334

Table 1.5: Comparison of ANN model and Linear Regression

Bayesian Model of Brittany Alexander:

This model Used Bayesian algorithm to predict US presidential elections at a state level. The objective was to predict the winner at each state and then provide an overall prediction. Ultimately, the model exhibited a high accuracy (number of correct state predictions over total number of states) for three election cycles 2008, 2012 and 2016 as shown in the following table (1.6).

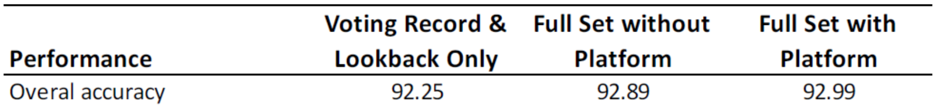


Race	Tested Model
2008 Accuracy	0.98039
2012 Accuracy	1
2016 Accuracy	0.88235
Average Accuracy	0.95425

Table 1.6: Percentages of states where winners were called correctly

Decision Tree model of Joseph Stoffa and other researchers:

This model used a decision tree algorithm to predict county-level voting for presidential elections. It attempted to predict the election outcomes of counties using independent variables like historical voting records, the number of past elections considered, socio-economic data and party platforms. The following table (1.7) illustrates the model's high accuracy for various combinations of independent variables.



Performance	Voting Record & Lookback Only	Full Set without Platform	Full Set with Platform
Overal accuracy	92.25	92.89	92.99

Table 1.7: Results of Decision Tree Model

Conclusion:

The modeling of U.S. elections began in the last decade when several economists and political scientists started using OLS regression to predict election outcomes. Ray C. Fair pioneered this approach of presidential election modeling, and others followed his lead, introducing new models with various sets of variables and predictors. Although these traditional modeling techniques yielded good results, researchers have recently begun using machine learning algorithms to achieve even greater forecasts. These advanced models have demonstrated significant progress and high accuracy, successfully predicting election winners with improved precision.

2. Methodology

Building on existing research previously discussed that focused on modeling US presidential elections, the methodology of this study Consists of numerous important steps. The review of previous works informed the selection of modeling techniques and variable definitions. Data was collected from different reliable sources, including historical election results, economic indicators, non-economic data, and political factors, ensuring a solid basis for the analysis. Both dependent and independent variables will be defined, with the dependent variable representing the election outcome, either as a continuous variable (percentage of votes) for regression models or as a binary variable (win/loss) for classification models. Independent variables included a range of economic and political indicators most of them identified through literature review. Multiple models will be constructed, including regression models such as Ordinary Least Squares (OLS) and Lasso regression to predict the percentage of votes, and classification models such as Logistic Regression, Naive Bayes, Decision Tree, and Artificial Neural Networks (ANN) to predict the binary outcome of winning or losing the election. Each model's performance will be assessed using appropriate metrics. This evaluation will determine the models' predictive accuracy and reliability. Based on the evaluation results, the models will be compared. Subsequently, they will be used to generate predictions for the 2024 US presidential election. By applying these methods, this study aims to contribute valuable insights for understanding and forecasting US presidential election outcomes.

While conducting this study, it is crucial to mention the assumption related to the bipartisan system of the United States. Despite the existence of other parties and candidates, the actual race will be between the Democratic nominee and the Republican nominee. When one side loses, the other party inevitably wins. For over 150 years, things have been that way. Independent candidates barely get any votes from the Electoral College. Almost all the studies mentioned in this report used this assumption.

2.1 Data collection

In order to gather data for this research, a variety of reliable sources were consulted, including official websites of American government and American organizations. Notably, valuable data related to the US elections were retrieved from the Federal Election Commission (FEC) website, which serves as the primary authority on electoral processes within the United States. Additionally, data was accessed from the International Monetary Fund's and International Financial Statistics (IMF/IFS) database, providing essential economic indicators for the analysis. Statista.com also was used as another crucial resource, offering plenty of statistical data and reports across various economic and political domains. For authoritative information on labor statistics, the U.S. Bureau of Labor Statistics (BLS) website was consulted. To search for historical presidential data and documents, resources such as The Presidency Project and the Presidency of the United States website hosted by the University of California, Santa Barbara (presidency.ucsb.edu) were utilized. Furthermore, the research utilized trusted sources such as Encyclopedia Britannica and Encyclopedia Wikipedia, ensuring the access to accurate information. The data was carefully cleaned and processed to make sure it was accurate and appropriate for analysis. The dataset covers 30 election cycles from 1904 to 2020, and all other collected data relate to this period. The data collected will be attached in Appendix B.

2.2 Variables Definition

2.2.1 Dependent variable

When trying to model the US presidential election, the primary focus of this project is to predict the outcome of the incumbent party. More precisely, this research intend to predict the Vote share of the candidate that represents the winning party in the previous election cycle. As assumed by several researchers mentioned previously, if the incumbent president who won the previous election has ruled over a period of sustained economic growth and prosperity, He will be re-elected. Or if he is not running for re-election, his successor from the same party will be chosen. The economic conditions immediately preceding a presidential election have a strong predictive effect on the election outcomes. These conditions could impact the voting behavior because of the assumption that voters consider how their own economic health has been affected during the tenure of the current president. Therefore, they will vote for the incumbent candidate if they see his performance was favorable, and vote for the challenger candidate otherwise.

The dependent variable in this project will capture the vote share received by the representative of the incumbent party in the election. Some of the studies in this topic used the popular vote as the response variable in their econometric models, while others used the electoral vote. The project will follow the method of popular vote for regression models. And for classification models, Electoral vote will be used as it is the one who eventually decide the winner.

As this study will incorporate two different modeling techniques, regression and classification, the independent variable will be represented in two ways:

- In regression models, the independent variable will be the percentage of popular votes of the incumbent party. If this percentage exceeds 50% then the candidate wins and his the party stays in power. Otherwise, the challenger from the other party will be the next president
- In classification models, the electoral vote will be considered. More precisely, the percentage of members that voted for the party in power from the Electoral College total voters (538). The independent variable will be a binary outcome. This binary data will be obtained by transforming the incumbent party's electoral vote percentage models as follow:
    - If the incumbent party's vote share exceeds 50%, the variable will take 1, indicating a win.
    - Otherwise, the variable will take the value 0, indicating a loss.

Using this approach, we ensure that the dependent variable will interpret the win or the loss of the incumbent party in the presidential election.

2.2.2 Independent variables

After reviewing previous literature on this topic, numerous relevant predictor features were identified. Building on these models and their approaches, this research will consider all the features from different models that were proven to be significant factors explaining the election outcomes. The method of Sinha (2020) will be used to classify the variables into two categories: economic and non-economic. Therefore, the models of this project will use the following listed explanatory variables.

Economic variables:

- ***GDP growth rate:*** This variable illustrates the percentage change in USA real GDP per capita between the year preceding the election and the election year itself. Real GDP per capita was used to mitigate the impact of population growth and inflation adjustments. It is computed as follows :

$$\text{Per Capita Real GDP Growth Rate} = \text{Real GDP Growth Rate} - \text{Population Growth Rate}$$

Equation 1

Where:

$$\text{Real GDP Growth Rate} = \left(\frac{\text{Real GDP in Current Year} - \text{Real GDP in Previous Year}}{\text{Real GDP in Previous Year}}\right) * 100$$

Equation 2

$$\text{Population Growth Rate} = \left(\frac{\text{Population in Current Year} - \text{Population in Previous Year}}{\text{Population in Previous Year}}\right) * 100$$

Equation 3

This variable was derived from Ray Fair's model

- ***Inflation rate:*** The inflation rate of the United States during election years. It refers to the percentage change in the general price level of goods and services over on an annual basis. It quantifies the rate at which the purchasing power of a currency declines from the year that precedes the election to that specific year. It is computed as follows:

$$\text{Inflation Rate} = \left(\frac{\text{Consumer Price Index in Current Year} - \text{Consumer Price Index in Previous Year}}{\text{Consumer Price Index in Previous Year}}\right) * 100$$

Equation 4

This variable was derived from Ray Fair's model

- ***Unemployment rate:*** The unemployment rate of the United stated during election years. It is a measure that indicates the percentage of the labor force that is currently unemployed and actively seeking employment.

$$\text{Unemployment Rate} = \left(\frac{\text{Number of Unemployed Persons}}{\text{Labor Force}}\right) * 100$$

Equation 5

This variable was derived from Pankaj Sinha's model

- ***Exchange rate:*** The exchange rate of British pound to American dollar GBP/USD. It is the number of American dollars that one British pound can buy over the years of election. It reflects broader global economic trends. Changes in this exchange rate may signify shifts in international trade and investment flows. The US and UK share significant trade relations, so changes in the exchange rate may affect trade dynamics between the two countries. Economic policies related to trade could be a central point in presidential campaigns, making the exchange rate a potentially relevant predictor.

  This variable was derived from Pankaj Sinha's model

- ***Stock market performance:*** This variable measures the overall health and movement of the stock market in the United States using the percentage change of the S&P 500 index that tracks the performance of 500 large-cap American stocks.

  This variable was derived from Cuzán & Bundrick's model

Non-economic variables:

- ***Approval rating:*** This variable measures the average job approval rating of the President in the USA. Which means how much the American people are satisfied with the work done by the president in his tenure of 4 years before the election .It is assessed through surveys conducted by organizations like Gallup. It serves as an indicator of public sentiment towards the President's performance and leadership. For example if the approval rating of the president is 0.4, it means that 40% of the peoples approve of his work

  This variable was derived from Lewis-Beck and Rice's model

- ***Midterm election seats:*** The midterm election in the USA is a national election held halfway through a president's four-year term. During midterms, voters elect members of Congress, including senators and house representatives. This variable compute the change of number of seats of the incumbent president's party. Which means the gain or loss of seats in both senate and house by the party in power. For example if the incumbent president is from the Democratic party, and this party loses 10 seats in the midterm election, the variables take the value -10

  This variable was derived from Lewis-Beck and Rice's model

- ***Geopolitical rating:*** This variable summarizes the most significant geopolitical events that happened during the incumbent president term and could affect the voters' perspective. These events will be classified into 5 major categories. Therefore, the geopolitical rating will take values from 0 to 5. The Categories of events are :
  - *Wars and Conflicts*: refer to important military engagements in which the United States has been directly or indirectly involved. These could include wars, military interventions, or any significant armed conflicts.
  - *Economic Crises:* encompass severe crises in the financial and economic systems of the United States. These may include events such as stock market crashes, banking crises, recessions, or depressions.
  - *Terrorist Attacks:* Significant terrorist attacks in the United States and include those with a high number of deaths, typically at least 10 victims.
  - *Political Scandals:* These are the unethical behaviors by president or his government officials and harm his reputation. Political scandals can range from allegations of corruption and abuse of power to personal indiscretions or violations of the law.
  - *Disasters and Accidents:* include both natural and industrial incidents resulting in significant loss of life, typically involving at least 1000 deaths. May include earthquakes, hurricanes, pandemics, oil spills…

  The assessment of a president's term includes an evaluation of whether significant events across various categories occurred during their tenure. Each category is treated as a distinct binary variable. It take the value of 1 if at least one event of that category occurred during the president's term and 0 otherwise. To aggregate these assessments into a comprehensive rating, each category is assigned equal weight. Therefore, the presence of any event within a category contributes equally to the final rating. The final rating, ranging from 0 to 5, reflects the cumulative occurrence of significant events across all categories

This variable was derived from Douglas Hibbs's model

- ***Campaign Spending Index:*** Campaign spending plays a crucial role in a presidential candidate's bid for office, as it directly influences their ability to reach voters and convey their message. Funding and advertising campaigns are regarded as essential for competitiveness in presidential elections. In assessing campaign spending's impact on electoral outcomes, especially when precise spending numbers are not very accurate across all election cycles, a dummy variable is introduced. It takes the value of 1 if the campaign expenditure of the candidate from the incumbent party (winning party in the previous election) surpasses that of the challenger, and 0 otherwise.

  This variable was derived from Pankaj Sinha's model

- ***Running for reelection:*** A dummy variable indicating whether the incumbent president is running for reelection or not in the next election cycle. It takes the value of 1 if the incumbent president is running for reelection and 0 if they are not.

  This variable was derived from Ray Fair's model

- ***Period in power:*** A dummy variable indicating whether the incumbent party was in power for two or more presidential terms. It takes 1 if the last two election cycles were won by the same party and 0 otherwise. For example if the democratic candidates won in 2008 and 2012, the variable will take the value 1 in 2016 election.

  This variable was derived from Alan Abramowitz's model

2.3 Regression models

2.3.1 OLS regression model

After defining the variables, An Ordinary Least Squares regression model will developed to explain the relationship between the presidential election outcomes and the independent variables discussed previously. This regression method was widely used in this area of study as mentioned previously. Ordinary Least Squares (OLS) regression is a statistical method used to estimate the relationship between a number of independent variables and a dependent variable. It does this through fitting a linear equation to the observations. The least squares approach seeks to minimize the sum of squared residuals and offers the best linear unbiased estimators (BLUE) of the coefficients, with assumptions of the model satisfying linearity, independence, homoscedasticity and normality of residuals.

The linear regression model will be fitted to the entire dataset, which contains 30 observations from presidential election cycles from 1904 to 2020. The OLS equation model is presented as follows:

$$Y_t = \beta 0 + \beta 1 * GGR_t + \beta 2 * IR_t + \beta 3 * UR_t + \beta 4 * ER_t + \beta 5 * SMP_t + \beta 6 * AR_t + \beta 7 * MES_t + \beta 8 * GR_t + \beta 9 * CSI_t + \beta 10 * RFR_t + \beta 11 * PIP_t + \varepsilon_t$$

Equation 6: Equation of the OLS model

Where:

- Y_t : The percentage of incumbent party's popular vote for time t
- $\beta 0$: The intercept
- $\beta 1$, $\beta 2$, $\beta 3$, … , $\beta 11$: Correlation coefficients of the independent variables
- GGR_t: GDP growth rate for time t
- IR_t: Inflation Rate for time t
- UR_t: Unemployment Rate for time t
- ER_t: Exchange Rate for time t
- SMP_t: Stock Market Performance for time t
- AR_t: Approval Rating for time t
- MES_t: Midterm Election Seats for time t
- GR_t: Geopolitical Rating for time t
- CSI_t: Campaign Spending Index for time t
- RFR_t: Running for Reelection for time t
- PIP_t: Period in Power for time t
- ε_t: Error term for time t

After developing the OLS model using the appropriate software, a backward elimination technique will be employed. This technique consists of selecting the features that build the model with the best performance. The results will be displayed and the model will be evaluated using various metrics, and tested for Multicollinearity, Heteroscedasticity and Autocorrelation.

2.3.2 LASSO regression model

LASSO regression method was used by Pankaj Sinha and other researchers (2020) when they wanted to make use of this machine learning algorithm in order to improve their predictions. They constructed an OLS model and they wanted to refine it using this algorithm. Following their method, this study will construct a LASSO regression model aiming to enhance the performance of the linear regression model.

Lasso regression (Least Absolute Shrinkage and Selection Operator regression) is a type of linear regression that includes a regularization term in the objective function. This regularization term is the sum of the absolute values of the coefficients multiplied by a tuning parameter (lambda). The purpose of using lasso regression is to prevent overfitting by shrinking some coefficients to zero. It is like performing variable selection to simplify the model. This algorithm is useful when dealing with a large number of predictors.

Lasso regression works by modifying the ordinary least squares (OLS) regression objective function and it includes a penalty term that reduce the size of some regression coefficients. The methodology of LASSO consists of solving the following optimization problem in order to get the Lasso vector of coefficients:

$$\beta^{LASSO} = \arg\ min_{\beta}\ \{\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j * x_{ij}\right)^2 + \lambda * \sum_{j=1}^{p}\left|\beta_j\right|\}$$

Equation 7: Optimization algorithm for Lasso Regression

Where:

y_i: The values of the dependent variable

x_{ij}: The values of the independent variables

β_j: The coefficients of the independent variables

λ: The regularization term

LASSO algorithm uses the objective function of an OLS and combine the residual sum of squares with the lasso penalty term, which is the sum of the absolute values of the coefficients scaled by λ.
The choice of λ is crucial, an optimal value should be determined because the higher the value of λ, the more coefficients are shrunk toward zero. If λ is zero, Lasso regression reduces to OLS regression.

The lasso regression acts as a variable selection technique, excluding variables seen as not relevant by shrinking their coefficients to zero. This algorithm will be applied and compared to the OLS model to determine if it provides any improvement.

2.4 Evaluation metrics for regression models

R-squared (R^2):

R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It indicates how well the independent variables explain the variability of the dependent variable.

$$R^2 = 1 - \frac{SSR}{SST}$$

Equation 8

Adjusted R-squared:

Adjusted R-squared adjusts the R-squared value by penalizing the addition of irrelevant independent variables to the model.

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2) * (n - 1)}{n - k - 1}$$

Equation 9

Root Mean Squared Error (RMSE):

RMSE measures the average magnitude of the residuals (the differences between predicted and actual values) in the model. It provides a measure of the model's accuracy in predicting the dependent variable.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$$

Equation 10

Mean Absolute Error (MAE):

MAE measures the average absolute difference between the predicted and actual values. It provides a measure of the model's performance, as it is less sensitive to outliers compared to RMSE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|$$

Equation 11

Where:

- SSR: the sum of squared residuals, sum of squared differences between the predicted and actual values
- SST: is the total sum of squares, sum of squared differences between the actual values and the mean of the dependent variable
- n: is the number of observations
- k: is the number of independent variables in the model
- y_i: is the actual value of the dependent variable for observation i
- $\widehat{y_i}$: is the predicted value of the dependent variable for observation i

2.5 Classification models

For the classification models, the independent variable will be all of the explanatory variables in the dataset. While the dependent variable will be transformed into a binary outcome. This transformation will be based on the concept of win or loss. If the incumbent party's percentage of electoral votes exceeds 50%, indicating a victory, the variable will be assigned a value of 1. Otherwise, it will be assigned a value of 0.

$$y = \begin{cases} 1 : if\ the\ percentage\ of\ Incumbent\ party's\ electoral\ votes\ exceeds\ 50\% \\ 0 :\ if\ the\ percentage\ of\ Incumbent\ party's\ electoral\ votes\ is\ below\ 50\% \end{cases}$$

The classification models will be trained and tested by splitting the dataset into training and testing sets with proportions of 0.6 and 0.4. The training set will include data for 18 election cycles from 1904 to 1972, while the testing set will consist of data for the subsequent 12 cycles from 1976 to 2020.

2.5.1 Artificial Neural Network model (ANN)

Artificial neural network (ANN) is a relatively new machine learning tool that has been widely used for forecasting in several fields. ANN are computational models inspired by the structure and function of biological neural networks in the human brain. They consist of interconnected nodes called neurons that are organized in layers. In classification tasks, ANNs can be used to learn complex patterns and relationships in data. The structure of ANN consists of an input layer, one or more hidden layers, and an output layer as shown in figure (2.1). Neurons are connected to neurons in the adjacent layers through weighted connections. The independent variables from the training set are provided as input to the ANN model. Also, the dependent variable of the training set is used for training the model. It represents the binary outcome to be predicted. The model learns from the input-output relationship in the training data through the process of forward propagation (passing input data through the network to make predictions) and backward propagation (adjusting weights based on prediction errors). This is done by adjusting the weights to minimize the loss function which measures the difference between predicted and actual outputs. Following Mohammad Zolghadr's approach (2018), An ANN model will be constructed.

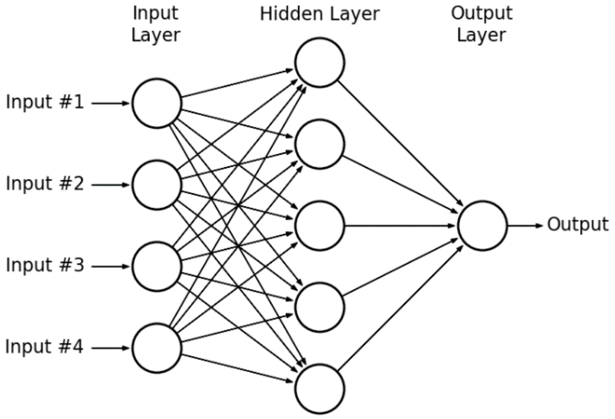


Figure 2.1: Structure of Artificial Neural Network Algorithm

2.5.2 Naive Bayes model

The Naïve Bayes classification model is a probabilistic machine learning model used for classification tasks. It is based on Bayes' Theorem and assumes that the independent variables are conditionally independent given the class label dependent variable. Despite the often unrealistic assumption of feature independence, Naive Bayes performs well in many practical applications in classification.

Through the training Phase, the algorithm calculates the prior Probabilities of each class in the training set. Then calculates Likelihood for each feature given each class. Subsequently the Bayes' Theorem uses the prior and likelihoods to compute the posterior probabilities of the classes given the set of features.

The classification of a given instance is done by computing the posterior probability for each class using the feature values of the instance. And then assign the class label with the highest posterior probability to the instance.

Bayes' Theorem:

$$P(C_k \mid x) = \frac{P(x \mid C_k) * P(C_k)}{P(x)}$$

Equation 12

The likelihood (assuming feature independence):

$$P(x \mid C_k) = \prod_{i=1}^{n} P(x_i \mid C_k)$$

Equation 13

Posterior Probability:

$$P(C_k \mid x) \propto P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)$$

Equation 14

Where:

- $P(C_k \mid x)$: The posterior probability of class C_k given features x
- $P(x \mid C_k)$: The likelihood of features x given class C_k
- $P(C_k)$: The prior probability of class C_k
- $P(x)$: The total probability of the features

As Brittany Alexander did (2019), this study will construct a Bayesian model to forecast US presidential election.

2.5.3 Logistic regression model

Logistic regression is a statistical method used for binary classification, in which the goal is to predict the probability of a binary outcome (dependent variable) based on one or more predictor variables (independent variables). It models the relationship between the dependent variable and the independent variables using the logistic function to ensure the predicted probabilities lie between 0 and 1.

The logistic regression model predicts the probability $\mathrm{P}(\ \mathrm{y} = 1 \mid \mathrm{X})$ that the dependent variable y equals 1 given the independent variables $\mathrm{X} = (X_1, X_2, \dots, X_3)$. The logistic function:

$$\mathrm{P}(\ \mathrm{y} = 1 \mid \mathrm{X}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2, \dots, + \beta_n X_n)}}$$

Equation 15

For the prediction, when a given instance with predictors X, the probability of the outcome y=1 is computed using the logistic function. A threshold (commonly 0.5) is applied to convert probabilities into binary predictions:

$$\hat{y} = \begin{cases} 1 : if\ \mathrm{P}(\ \mathrm{y} = 1 \mid \mathrm{X}) \geq 0.5 \\ 0 : \ if\ \mathrm{P}(\ \mathrm{y} = 1 \mid \mathrm{X}) < 0.5 \end{cases}$$

Following Nate Silver’s methodology, who used logistic regression in his approach, this study will develop a logistic regression model to predict the US presidential election.

2.5.4 Decision Tree model

A decision tree is a supervised learning algorithm used for classification tasks. It models decisions and their possible consequences as a graph in the form of a tree of decisions. Each internal node of the tree represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label for classification. The tree Construction starts with the root node that uses the entire dataset and then perform the splitting process. At each node, the dataset is split into subsets based on the value of an attribute. The goal is to find the attribute and the threshold that best separate the data into homogeneous subsets. Criteria for splitting includes Gini Impurity (Measures the frequency of different classes at a node) and Entropy (Used in information gain to measure the disorder or impurity). The splitting process is applied recursively to each subset, creating branches and further nodes. And it stops when all instances in a node belong to the same class and no further splits can improve the purity of the node. Subsequently, Tree Pruning is used to reduce the size of the tree and prevent overfitting. It involves removing branches that have little importance.

For predicting a given instance, the decision tree is traversed from the root node to a leaf node by following the decisions at each node based on the attribute values of the instance. The class label at the leaf node is assigned as the prediction of the instance.

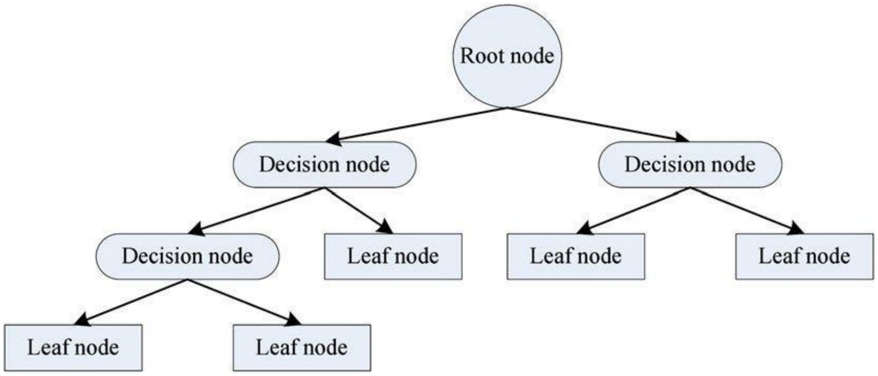


Figure 2.2: Structure of the Decision Tree algorithm

Gini Impurity:

$$\mathrm{Gini(D)} = 1 - \sum_{i=1}^{C} p_i^2$$

Equation 16

Entropy (Information Gain):

$$\mathrm{Entropy(D)} = -\sum_{i=1}^{C} p_i * \log(p_i)$$

Equation 17

Where p_i is the proportion of instances of class I in dataset D, and C is the number of classes.

Following the steps of researcher Joseph Stoffa (2018), this study will develop a decision tree model to forecast US presidential elections.

2.6 Evaluation metrics for classification models

Accuracy:

The ratio of correctly predicted instances to the total instances:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 18

Precision:

The ratio of correctly predicted positive instances to the total predicted positives:

$$Precision = \frac{TP}{TP + FP}$$

Equation 19

Recall: The ratio of correctly predicted positive instances to all actual positives:

$$Recall = \frac{TP}{TP + FN}$$

Equation 20

F1 Score:

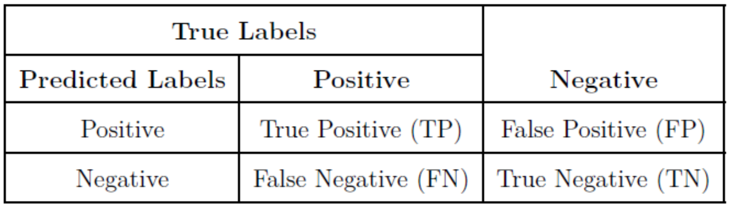
The harmonic mean of precision and recall:

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Equation 21

Confusion Matrix:

A table used to describe the performance of the classification model:



True Labels		Negative
Predicted Labels	**Positive**	
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

Table 2.1: Confusion Matrix structure

Where:

- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN: False Negatives

3. Implementation

In this section, the project's processes will be implemented and empirical results will be presented. First, a descriptive analysis will be conducted to further understand the variables and the data. Then, the previously discussed models will be implemented to display their results and they will be assessed using evaluation metrics. Finally, the models will be used to predict the outcome of the 2024 US presidential election.

3.1 Exploratory Data Analysis

3.1.1 Descriptive Statistics

This section of the study aims to understand the data and identify its patterns. Through examination and visualization of this data, this study intend to gain valuable insights of the factors shaping US presidential election and the relationships that could influence election outcomes. The dataset spans from 1904 to 2020, composed of economic and non-economic data. The table 3.1 shows the descriptive statistics for the variables.

Variable	**Popular percentage**	**Inflation rate**	**GDP growth rate**	**Exchange rate**	**Unemployment rate**	**Stock market performance**
Mean	0,4931	2,9	0,016946	2,960	6,723	0,082017
Standard Error	0,015496	0,819	0,009573	0,239	0,809	0,031
Median	0,495	2,1	0,023640	2,79	5,45	0,1066
Range	0,379	25,5	0,273202	3,69	22,4	0,7702
Minimum	0,232	-9,9	-0,151060	1,28	1,2	-0,3849
Maximum	0,611	15,6	0,122142	4,97	23,6	0,3853
Variable	**Campaign Spending Index**	**Approval rating**	**Midterm election seats**	**Geopolitical rating**	**Running for reelection**	**Period in power**
Mean	0,667	0,544633	-33,6	2,033333	0,6	0,6
Standard Error	0,088	0,017	5,227	0,217	0,091	0,0901
Median	1	0,5425	-30,5	2	1	1
Range	1	0,377	106	4	1	1
Minimum	0	0,365	-88	0	0	0
Maximum	1	0,742	18	4	1	1

Table 3.1: Descriptive Statistics

The following figures (3.1 and 3.2) present visualizations of the target variable and some key predictors.

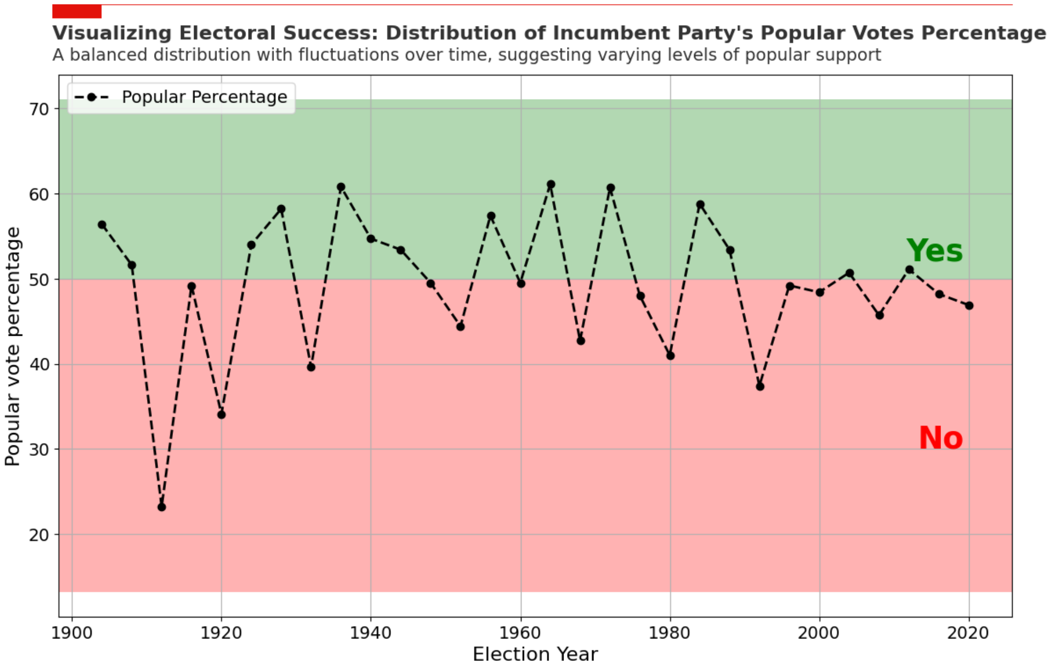


Figure 3.1: Popular percentage of the incumbent party during election years

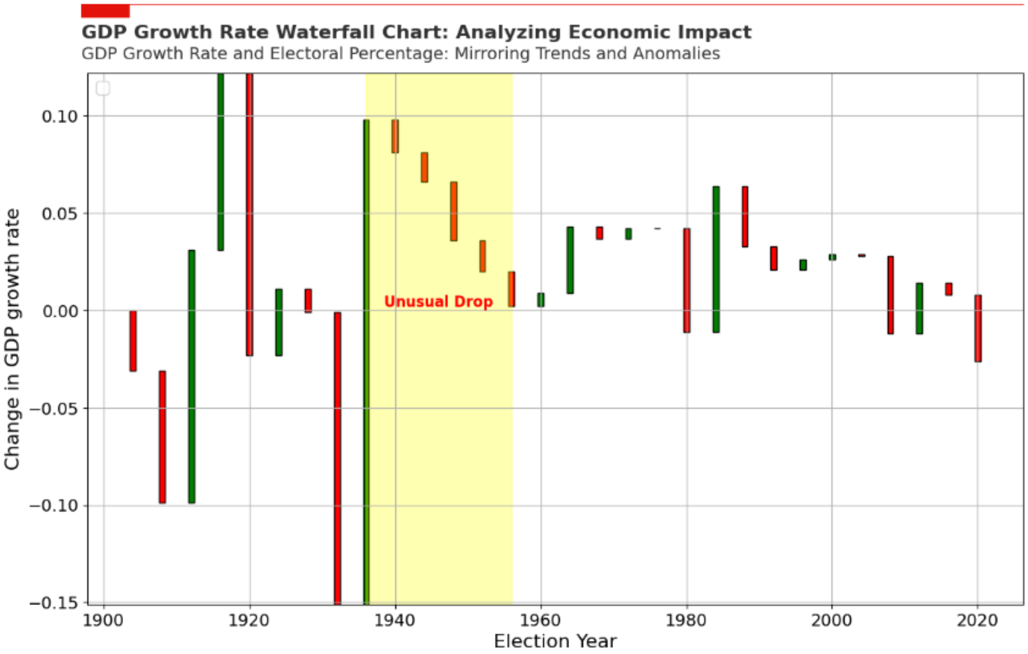


Figure 3.2: Waterfall chart of GDP growth rate during election years

A notable observation from this chart is the significant drop in GDP growth rate between 1936 and 1952. This period coincides with the Great Depression and World War II, which likely influenced the GDP.

The rest of the visualizations are attached in appendix A

3.1.2 Correlation Matrix

The following table (3.2) presents the correlation matrix for the independent variables considered in the research.

	IR	*GGR*	*ER*	*UR*	*SMP*	*CSI*	*AR*	*MES*	*GR*	*RFR*	*PIP*
Inflation rate	1										
GDP growth rate	0,3903	1									
Exchange Rate	-0,153	-0,0127	1								
Unemployment rate	-0,447	-0,3171	0,1226	1							
stock market performance	-0,152	-0,006	0,2222	-0,1362	1						
Campaign Spending Index	0,0208	0,2767	0,2447	-0,0946	0,27977	1					
Approval rating	-0,208	-0,0008	0,3598	0,1506	0,24327	0,458	1				
Midterm election seats	0,0859	-0,0255	-0,123	-0,0736	0,24168	0,186	0,501	1			
Geopolitical rating	0,0407	-0,0263	-0,56	-0,1607	-0,3993	-0,161	-0,328	-0,1585	1		
Running for reelection	-0,085	0,206	-0,044	0,3475	0,11495	0,144	0,1645	0,0382	-0,21	1	
Period in power	-0,114	-0,2769	0,1766	-0,047	-0,2758	-0,289	-0,188	-0,1189	0,0233	-0,5278	1

Table 3.2: Correlation Matrix of independent variables

The Pearson correlation coefficient was used to test the statistical correlation between each pair of independent variables in the dataset. A higher correlation coefficient indicates a greater degree of dependence between the variables. The Correlation Matrix shows that there is moderate correlation between the independent variables ranging from -0,56 to 0.501. None of the correlation coefficients are close to -1 or 1 with indicate a lack of strong correlation. While there is some degree of correlation, it is not strong enough to imply a very high dependence between any pairs of variables.

To explore the multicollinearity between independent variables, a test of Variance Inflation Factor (VIF) was performed leading to the following results in table (3.3).

Variable	VIF
Inflation rate	1.694293
GDP growth rate	1.529889
Exchange Rate	2.200578
Unemployment rate	1.828257
stock market performance	1.721230
Campaign Spending Index	1.542855
Approval rating	2.364209
Midterm election seats	1.867885
Geopolitical rating	1.987557
Running for reelection	1.820563
Period in power	1.894632

Table 3.3 Variance Inflation Factor (VIF) test of Independent variable

All VIF values are significantly below 10. Therefore, the independent variables do not have a multicollinearity problem.

3.2 Regression models implementation

3.2.1 OLS regression model results

After running the developed OLS regression model with the dependent and independent variables, the following results are displayed in table (3.4) and table (3.5).

Regression statistics	
R-squared	0.618
Adjusted R-squared	0.384
Standard Error	0,067
F-statistic	0.0325
RMSE	0.0515
MAE	0.0407
Observations	30

Table 3.4: OLS regression statistics

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0,606767172	0,135797453	4,468177851	0,000297149
Inflation rate	-0,00859988	0,003587766	-2,397001191	0,027596168
GDP growth rate	0,566121307	0,291762916	1,940347028	0,06816722
Exchange Rate	-0,01169707	0,014036654	-0,833323254	0,415583778
Unemployment rate	0,000892228	0,003772046	0,236536734	0,815687046
stock market performance	0,032194299	0,095822761	0,335977579	0,740772467
Campaign Spending Index	0,064814573	0,032043285	2,022719372	0,058216109
Approval rating	-0,000644923	0,204932724	-0,003146999	0,997523678
Midterm election seats	0,000515096	0,000590419	0,872425661	0,394465262
Geopolitical rating	-0,01423012	0,01467146	-0,969918482	0,344940662
Running for reelection	-0,055889206	0,033493885	-1,668639109	0,112489701
Period in power	-0,059039901	0,034168434	-1,727907732	0,101121844

Table 3.5: OLS regression coefficients

To improve the results of this regression model, a backward elimination technique is employed in order to improve the adjusted R square by eliminating the least significant variables. The following table 3.6 traces this process.

Model	Evaluation metrics	p-values of significant variables (5% level)	Variable eliminated (Highest p-value)

Y_t=β0 + β1*GGR_t+ β2*IR_t + β3*UR_t + β4*ER_t + β5*SMP_t +β6*AR_t + β7* MES_t + β8*GR_t + β9*CSI_t + β10*RFR_t + β11*PIP_t	Adjusted R^2= 0.384
	RMSE= 0.0515
	MAE= 0.0407
		IR_t = 0.028	AR_t
			(0.998)

Y_t=β0 + β1*GGR_t+ β2*IR_t + β3*UR_t + β4*ER_t + β5*SMP_t + β7*MES_t + β8*GR_t+ β9*CSI_t + β10*RFR_t + β11*PIP_t	Adjusted R^2= 0.417
	RMSE= 0.0516
	MAE= 0.0408
		IR_t = 0.020
		CSI_t= 0.041
			UR_t
			(0.811)

Y_t=β0 + β1*GGR_t+ β2*IR_t + β4*ER_t + β5*SMP_t + β7*MES_t + β8*GR_t + β9*CSI_t + β10*RFR_t + β11*PIP_t	Adjusted R^2= 0.444
	RMSE= 0.0517
	MAE= 0.0402
		IR_t = 0.008
		CSI_t= 0.037
			SMP_t
			(0.775)

Y_t=β0 + β1*GGR_t+ β2*IR_t + β4*ER_t + β7*MES_t + β8*GR_t+ β9*CSI_t + β10* RFR_t + β11*PIP_t	Adjusted R^2= 0.468
	RMSE= 0.0518
	MAE= 0.0406
		IR_t = 0.005
		GGR_t= 0.048
		CSI_t= 0.029
		PIP_t= 0.046
			ER_t
			(0.356)

Y_t=β0 + β1*GGR_t+ β2*IR_t + β7*MES_t + β8*GR_t+ β9*CSI_t + β10*RFR_t + β11* PIP_t	Adjusted R^2= 0.470
	RMSE= 0.0526
	MAE= 0.0400
		IR_t = 0.006
		GGR_t= 0.048
		CSI_t= 0.041
		PIP_t= 0.029
			GR_t
			(0.394)

Y_t=β0 + β1*GGR_t+ β2*IR_t + β7*MES_t + β9*CSI_t + β10*RFR_t + β11*PIP_t	Adjusted R^2= 0.477
	RMSE= 0.0538
	MAE= 0.0406
		IR_t = 0.006
		GGR_t= 0.048
		CSI_t= 0.029
		PIP_t= 0.035
			RFR_t
			(0.116)

Y_t=β0 + β1*GGR_t+ β2*IR_t + β7*MES_t + β9*CSI_t + β11*PIP_t	Adjusted R^2= 0.440
	RMSE= 0.0568
	MAE= 0.0412
		IR_t = 0.014
		CSI_t= 0.028
			PIP_t
			(0.127)

Table 3.6: OLS variables selection

The last iteration generated a lower adjusted R-squared and higher RMSE and MAE than the previous one. This iteration didn't improve the performance of the model. Therefore, the process was stopped, and the retained model contains the following variables: GDP growth rate, inflation rate, midterm election seats, campaign spending index, running for reelection and period in power. The following equation represents the final model:

$$Y_t = \beta_0 + \beta_1 * GGR_t + \beta_2 * IR_t + \beta_3 * MES_t + \beta_4 * CSI_t + \beta_5 * RFR_t + \beta_6 * PIP_t + \varepsilon_t$$

Equation 22: Equation of final OLS model

The final model reveals the results presented in table (3.7). And the table (3.8) presents the coefficients and p-values of the independent variables.

Regression statistics	
R-squared	0.585
Adjusted R-squared	0.477
Standard Error	0,061
F-statistic	0.0013
RMSE	0.0538
MAE	0.0406
Observations	30

Table 3.7: OLS final model regression statistics

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0,557942924	0,042932481	12,99582309	4,43073E-12
Inflation rate	-0,008734397	0,002866403	-3,047163341	0,005719104
GDP growth rate	0,537776542	0,25703987	2,092191152	0,047653914
Campaign Spending Index	0,060803873	0,026071966	2,332155312	0,028808448
Midterm election seats	0,000705122	0,000411581	1,713204353	0,100124542
Running for reelection	-0,045305426	0,027704694	-1,635297816	0,115600202
Period in power	-0,063812303	0,028497306	-2,23923982	0,035104238

Table 3.8: OLS final regression coefficients

- Adjusted R-squared (0.477): indicates that approximately the model explains about 47.7% of the variance in the dependent variable in the model. This demonstrates a moderate level of explanatory power between the predictors and the outcome variable.
- F-statistic (0.0013): The F-statistic test of the overall significance of the regression model shows A very low p-value (<0.05) indicating that the model is globally statistically significant.
- Root Mean Squared Error (0.0538): RMSE shows a value of 0.0538 suggesting that the model's predictions varies from the actual values by 0.0538 units on average.
- Mean Absolute Error (0.0406): MAE shows a value of 0.0406 indicating that the predictions deviate from the actual values by about 0.0406 units on average.

The following figure (3.3) shows how the predicted values of the OLS model vary from the actual values of the popular percentage:

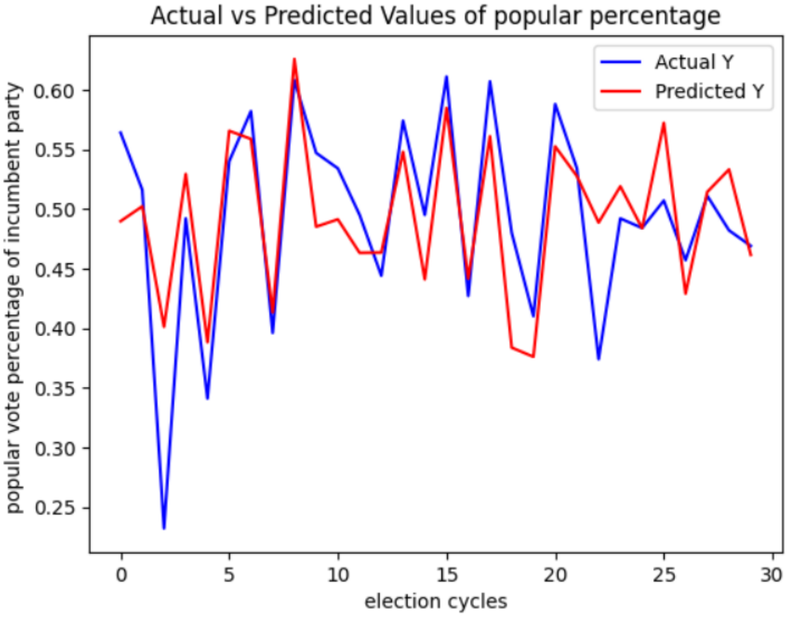


Figure 3.3: Actual vs Predicted values of popular vote percentage of the incumbent party

Significance of variables:

The following table (3.9) presents the hypothesis testing for the significance of the variables included in the final model. The significance level used is 5%.

The variable	Hypothesis	Results
GDP growth rate	H0: GGR_t is not significant H1: GGR_t is statistically significant	p-value = 0.048 < 0.05 ⇒ Reject H0
Inflation rate	H0: IR_t is not significant H1: IR_t is statistically significant	p-value = 0.006 < 0.05 ⇒ Reject H0
Midterm election seats	H0: MES_t is not significant H1: MES_t is statistically significant	p-value = 0.100 > 0.05 ⇒ Fail to reject H0
Campaign spending index	H0: CSI_t is not significant H1: CSI_t is statistically significant	p-value = 0.029 < 0.05 ⇒ Reject H0
Running for reelection	H0: RFR_t is not significant H1: RFR_t is statistically significant	p-value = 0.116 > 0.05 ⇒ Fail to reject H0
Period in power	H0: PIP_t is not significant H1: PIP_t is statistically significant	p-value = 0.035 < 0.05 ⇒ Reject H0

Table 3.9: Significance testing of the variables

At a 5% level of significance, GDP growth rate, Inflation rate, Campaign spending index and Period in power are significant. While Midterm election seats and Running for reelection are not significant at 5% level. They are significant only at 15% level.

Checking for Heteroscedasticity:

When performing a Breusch-Pagan/Cook-Weisberg test for the proposed OLS model, the p-value generated was 0.089 which is greater than 0.05. Consequently the model does not suffer from a Heteroscedasticity problem.

Checking for Autocorrelation:

The proposed OLS model exhibits a Durbin-Watson Value of 1.471. Given that the model has 30 observations and 6 predictors, the lower bound (dL) is around 0.95 and the upper bound (dU) is around 1.74. The Durbin-Watson Value of 1.471 lies between these two values, meaning that the test result is inconclusive. An inconclusive result suggests that while there is some indication of positive autocorrelation, it is not strong enough to be definitive.

3.2.2 LASSO regression model results

Before running the LASSO regression model with the dependent and independent variables, an optimal value of regularization parameter lambda should be determined. With executing a code the optimal value was found to be $\lambda = 0.0002$

The results of the Lasso regression model using the optimal regularization parameter are presented in the table (3.10) below:

Regression statistics	
R-squared	0.614
Adjusted R-squared	0.341
RMSE	0.0519
MAE	0.0399

Table 3.10: LASSO regression statistics

The following table (3.11) shows the estimated coefficients of the independent variables by the lasso regression model:

Variable	β
Inflation rate	-0.008342
GDP growth rate	0.446547
Exchange Rate	-0.010945
Unemployment rate	0.000237
stock market performance	0.017688
Campaign Spending Index	0.066658
Approval rating	0
Midterm election seats	0.000510
Geopolitical rating	-0.014416
Running for reelection	-0.050473
Period in power	-0.060042

Table 3.11: LASSO regression coefficients

The Lasso regression model performed variable selection on the model by eliminating the approval rating variable by shrinking its coefficient to 0. Other coefficients were also shrunk.

Even with an optimal value of lambda, the Lasso regression did not enhance the performance of the OLS model, as it reduced both the R-squared and Adjusted R-squared values. Additionally, the RMSE increased, resulting in a larger error margin. Contrary to its effect on Sinha’s model (2020), Lasso regression does not appear to be a suitable option for this set of variables. This suggests that all predictors in this model are relevant and that there is a lack of multicollinearity, as Lasso regression tends to exclude variables who have multicollinearity issues.

3.3 Classification models implementation

After splitting the dataset into training and testing sets, four classification models were implemented and tested: ANN, Naïve Bayes, Logistic regression and Decision Tree. The following table (3.12) shows the prediction outcomes of these models in the testing phase:

Election cycle	Actual value	ANN	Naïve Bayes	Logistic regression	Decision tree
1976	0	0	0	0	0
1980	0	1	0	0	0
1984	1	1	1	1	1
1988	1	1	1	1	1
1992	0	0	1	1	1
1996	1	1	1	1	1
2000	0	0	0	0	0
2004	1	1	1	1	1
2008	0	0	0	0	0
2012	1	1	1	1	1
2016	0	0	0	1	1
2020	0	0	0	0	0

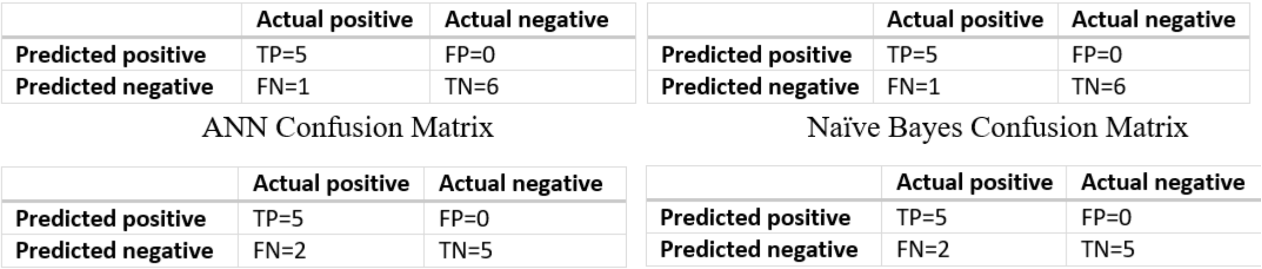
Table 3.12: Predicted outcomes of election using classification models

To evaluate the performance of these models during the testing phase, evaluation metrics were computed based on the predicted outcomes. The next table (3.13) shows the values of these evaluation metrics:

Model	Correct Predictions rate	Accuracy	Precision	F1 score
ANN	11/12	0.916	0.833	0.909
Naïve Bayes	11/12	0.916	0.833	0.909
Logistic regression	10/12	0.833	0.714	0.833
Decision tree	10/12	0.833	0.714	0.833

Table 3.13: Evaluation metrics for classification models

The Following Figure (3.4) provides the Confusion Matrix of each Model After the testing phase:



	Actual positive	Actual negative
Predicted positive	TP=5	FP=0
Predicted negative	FN=1	TN=6

ANN Confusion Matrix

	Actual positive	Actual negative
Predicted positive	TP=5	FP=0
Predicted negative	FN=1	TN=6

Naïve Bayes Confusion Matrix

	Actual positive	Actual negative
Predicted positive	TP=5	FP=0
Predicted negative	FN=2	TN=5

Logistic Regression confusion matrix

	Actual positive	Actual negative
Predicted positive	TP=5	FP=0
Predicted negative	FN=2	TN=5

Decision Tree Confusion Matrix

Figure 3.4: Confusion Matrices of classification Models

3.4 Predicting 2024 US presidential election

To forecast the results of the upcoming 2024 election, data was collected to identify the predictor values for this year. The identified values are presented in the following table (3.14).

Variable	2024 value
Inflation rate	3.3
GDP growth rate	0.011
Exchange Rate	1.28
Unemployment rate	3.9
stock market performance	0.123
Campaign Spending Index	0
Approval rating	0.40
Midterm election seats	-8
Geopolitical rating	3
Running for reelection	1
Period in power	0

Table 3.14: Data values for 2024 election

After running both regression and classification models on these values, the prediction results are presented as follows:

- OLS regression model: generated a value of 47.8, indicating that the incumbent party will lose, as the predicted percentage is below the 50%.
- LASSO regression model: generated a value of 47.6, indicating that the incumbent party will lose, as the predicted percentage is below the 50%.
- ANN model: generated an outcome of 0, indicating the incumbent party will lose
- Naïve Bayes model: generated an outcome of 0, indicating the incumbent party will lose
- Logistic Regression model: generated an outcome of 0, indicating the incumbent party will lose
- ANN model: generated an outcome of 0, indicating the incumbent party will lose
- Decision Tree model: generated an outcome of 0, indicating the incumbent party will lose

All developed models yielded the same prediction result, which is the loss of the incumbent party. According to these forecasts, the current president of the USA and Democratic candidate, Joe Biden, will lose to his challenger, the Republican candidate Donald Trump. As a result, the presidency will move from the Democratic Party to the Republican Party.

Conclusion

The purpose of this project was to study US presidential election modeling techniques and to utilize these methodologies to develop specialized forecasting models for predicting election outcomes. An extensive literature review was conducted to inform the project's modeling approaches. To identify the factors influencing voter behavior and election results, both regression analysis and classification methods were employed to construct predictive models for US presidential elections.

After developing these models, the relationship between election outcomes and various economic and political factors was found to be significant, confirming the findings of previous researchers. The variables used, encompassing both economic and non-economic data, demonstrated some degree of predictive power concerning election outcomes.

Firstly, the regression analysis incorporated several potential features, inspired by prior research, to determine the most significant predictors. Among the economic variables, GDP growth rate and inflation rate during election years were found to be particularly significant and relevant for predicting election results. Additionally, from the non-economic variables, the campaign spending index proved to be a significant predictor. This variable indicates whether the incumbent party's candidate has outspent their challenger in campaign spending. After constructing the final model and retaining the most significant variables, the results were acceptable, with a reasonably small margin of error. The model was then applied to predict the 2024 election outcome and it predicted the loss of the incumbent party democratic nominee Joe Biden to republican nominee Donald Trump by a percentage of 47.8%

Furthermore, the classification models employed machine learning algorithms such as Artificial Neural Networks (ANN), Naïve Bayes, Logistic Regression, and Decision Trees. Initially, these models were trained on the first 18 observations of the dataset, using all the independent variables with a transformed binary election outcome as the dependent variable. Subsequently, the models were tested on the remaining 12 observations. The results demonstrated high accuracy, with the ANN and Naïve Bayes models correctly predicting 11 out of 12 outcomes, and the Logistic Regression and Decision Tree models correctly predicting 10 out of 12. Ultimately, all four models predicted the loss of the incumbent party's Democratic nominee, Joe Biden, to the Republican nominee, Donald Trump, which aligns with the prediction of the regression model.

The findings suggest that both regression and classification approaches are viable options for modeling US presidential elections, as they provided reasonable and acceptable results. While classification models such as ANN, Naïve Bayes, Logistic Regression, and Decision Trees offered high accuracy in predicting the election winner, regression models have the advantage of providing an exact percentage that indicates the projected margin of victory. Both methods yielded the same prediction for the 2024 presidential election cycle: a loss for the Democratic Party to the Republican Party. The alignment of predictions across different models further validates the reliability of the methodologies used in this project.

Limitations

This research faced some limitations that should be noted. First, the economic data used covered the entire election year. Extracting precise economic data immediately before the election was not feasible, as accessing such specific data is challenging. This limitation may affect the accuracy of the models. Moreover, the data for predicting the 2024 election results was available up to the date of conducting this research and does not represent the entire election year.

Second, some variables used in previous models were not available for this study. These variables might be significant factors influencing election outcomes, and their omission could lead to an incomplete analysis.

Additionally, this research, like many previous studies, assumes that only two parties compete for the presidency, ignoring the impact of independent candidates and minor parties. Although these candidates typically receive a small percentage of the vote, their presence could influence the overall results.

Lastly, the models may not adequately account for unexpected events such as economic crises, natural disasters, or political scandals, which can significantly affect election outcomes. These events are difficult to predict and incorporate into the models, thus potentially impacting the accuracy of the predictions.

By identifying these limitations, the study highlights the areas in which more improvement and data collection may enhance both the accuracy and reliability of election result forecasts.

References

1. Abramowitz A. I. (1988). An Improved Model for Predicting the Outcomes of Presidential Elections. PS: Political Science and Politics, 21 4, 843-847
2. Abramowitz, A. (2012). Forecasting in a Polarized Era : The Time for Change Model and the 2012 Presidential Election. PS: Political Science & Politics, 45(4), Article 4. https://doi.org/10.1017/S104909651200087X
3. Alfred G. Cuzán. (2008). Forecasting U.S. Presidential Elections : A Brief Review (International Institute of Forecasters). 10, 29-34. The International Journal of Applied Forecasting.
4. Brittany Alexander. (2017). A Bayesian Model for the Prediction of United States Presidential Elections (Texas Tech University). Department of Mathematics and Statistics.
5. Campbell, J. E. (2008). Evaluating U.S. presidential election forecasts and forecasting equations. International Journal of Forecasting, 24(2), Article 2. https://doi.org/10.1016/j.ijforecast.2008.03.001
6. Cuzán, A. G., & Bundrick, C. M. (2009). Predicting Presidential Elections with Equally Weighted Regressors in Fair's Equation and the Fiscal Model. Political Analysis, 17(3), Article 3. https://doi.org/10.1093/pan/mpp008
7. Cuzán, A. G., Heggen R.J., & Bundrick C.M. (2000). Fiscal policy, economic conditions, and terms in office: simulating presidential election outcomes. In Proceedings of the World Congress of the Systems Sciences and ISSS International Society for the Systems Sciences, 44th Annual Meeting, July 16–20, Toronto, Canada.
8. Cuzán, A.G. & Bundrick, C.M. (2005). Deconstructing the 2004 presidential election forecasts: The fiscal model and the Campbell collection compared, PS: Political Science and Politics, 38(2), 255-262.
9. Erikson, R. S. (1989). Economic Conditions and the Presidential Vote. American Political Science Review, 83(2), Article 2. https://doi.org/10.2307/1962406
10. Erikson, R. S., and Wlezien, C. (1996). Of time and presidential election forecasts. PS: Political Science and politics, 31, 37-39
11. Fair, R. C. (1978). The effect of economic events on votes for president. Review of Economics and Statistics, 60, 159-173
12. Hibbs D. A. (2000). Bread and Peace voting in U.S. presidential elections. Public Choice, 104, 149–180.
13. Lewis-Beck, M. S. & Rice, T. W. (1982).Presidential Popularity and Presidential Vote. The Public Opinion Quarterly, 46 4, 534-537
14. Prechter, R. R., Goel, D., Parker, W. D., & Lampert, M. (2012). Social Mood, Stock Market Performance, and U.S. Presidential Elections : A Socionomic Perspective on Voting Results. SAGE Open, 2(4), Article 4. https://doi.org/10.1177/2158244012459194
15. Sinha, Pankaj and Verma, Aniket and Shah, Purav and Singh, & Jahnavi and Panwar. (2020). Prediction for the 2020 United States Presidential Election using Machine Learning Algorithm : Lasso Regression (Faculty of Management Studies).
16. Stoffa, J., Lisbona, R., Farrar, C., & Martos, M. (2018). Predicting How U.S. Counties will Vote in Presidential Elections through Analysis of Socio-Economic Factors, Voting Heuristics, and Party Platforms, 1(1), Article 1.
17. Zolghadr, M., Niaki, S. A. A., & Niaki, S. T. A. (2018). Modeling and forecasting US presidential election using learning algorithms. Journal of Industrial Engineering International, 14(3), Article 3. https://doi.org/10.1007/s40092-017-0238-2

Appendices

Appendix A

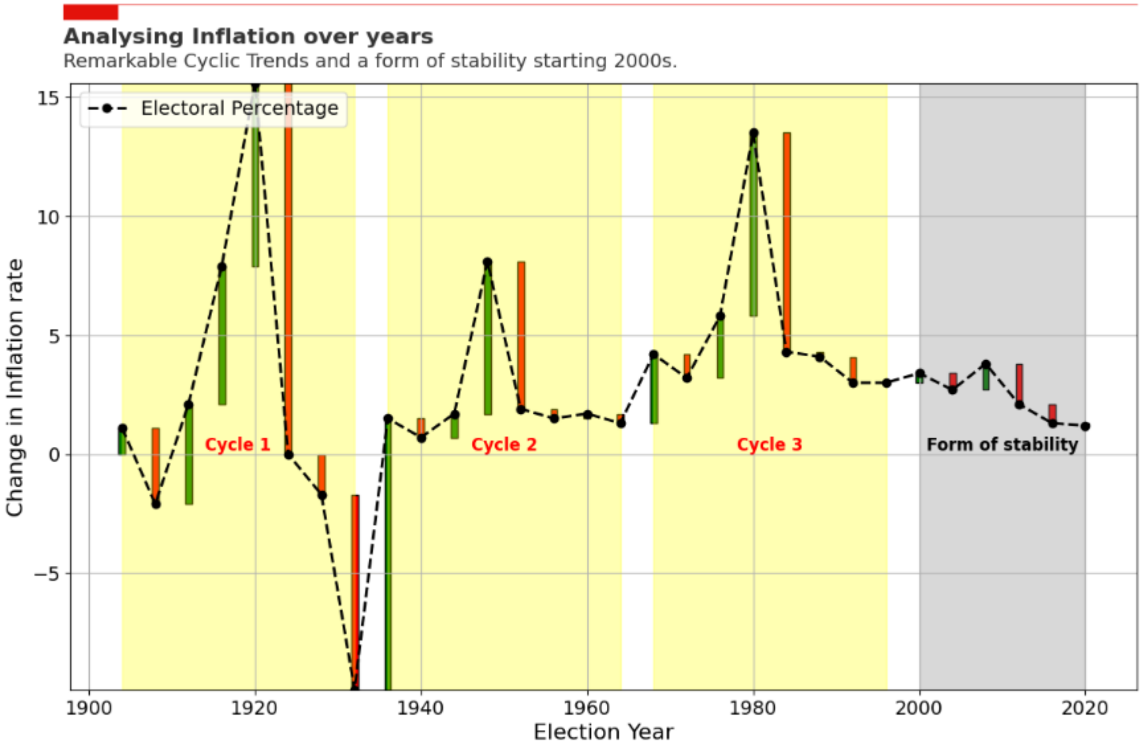


Figure A.1: Waterfall chart of inflation rates during election years

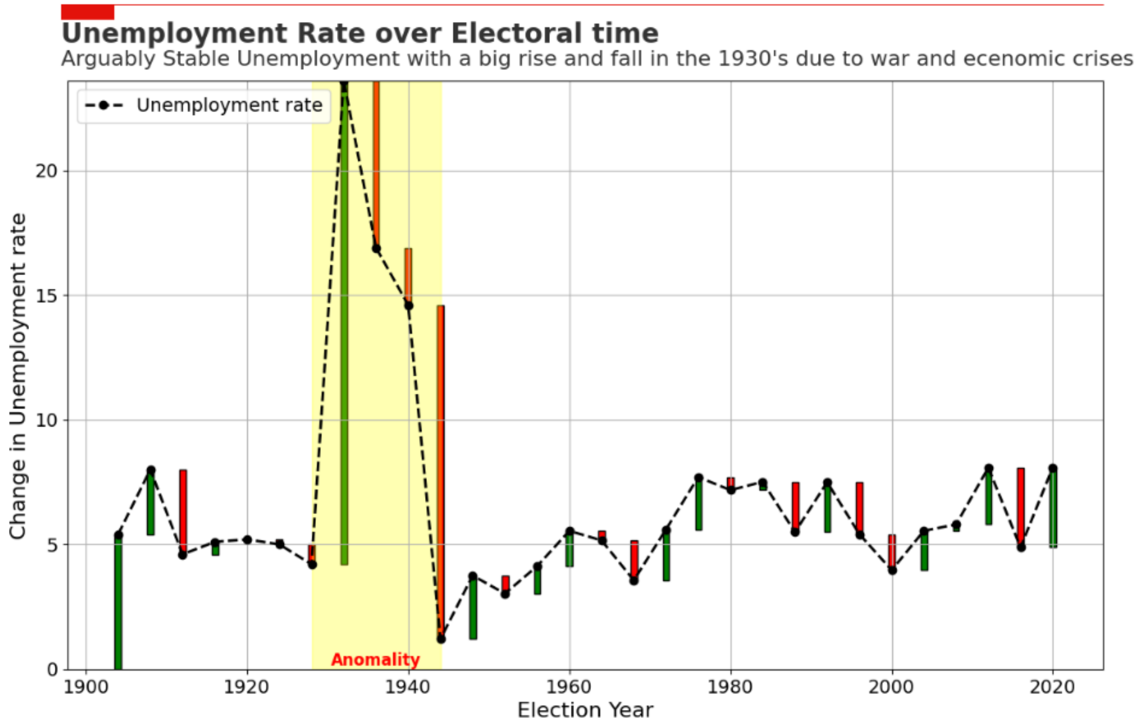


Figure A.2: Waterfall chart of Unemployment rates during election years

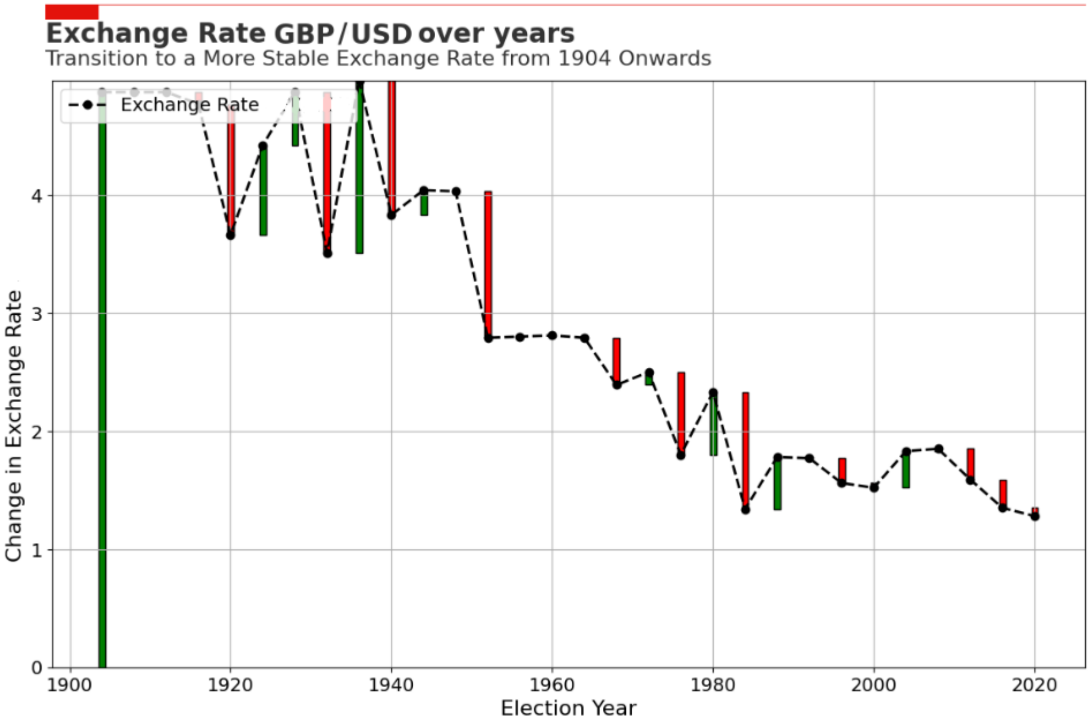


Figure A.3: Waterfall chart of GBP/USD Exchange Rates during election years

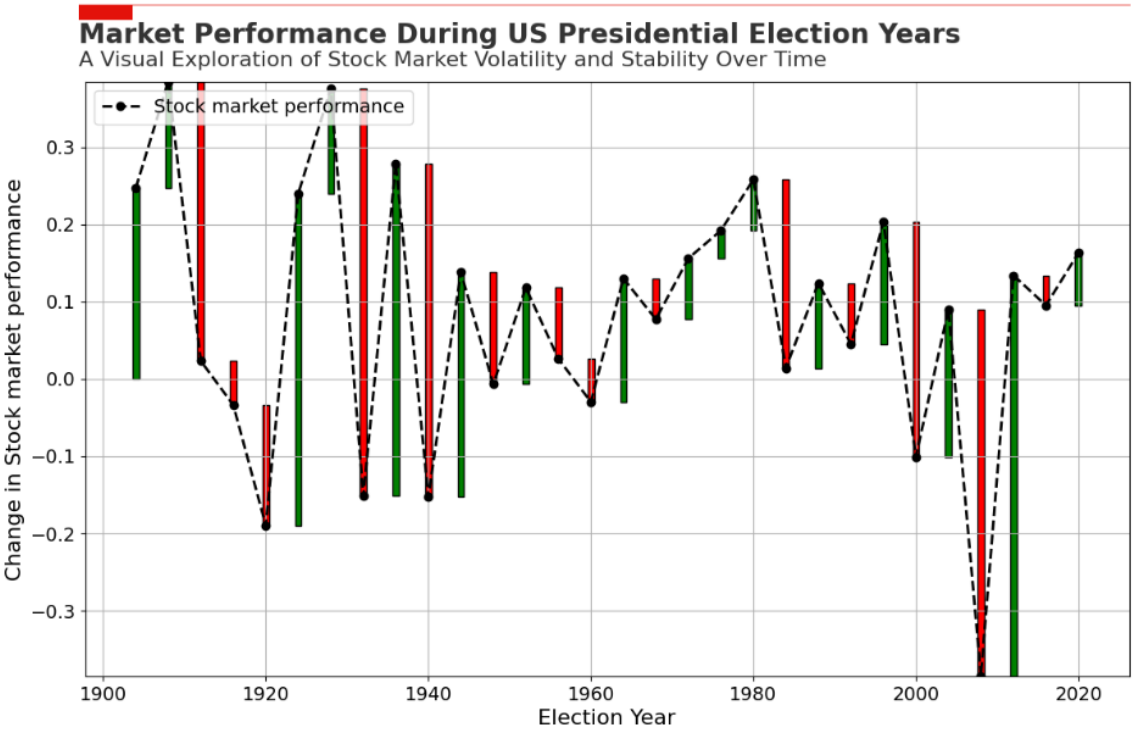


Figure A.4: Waterfall chart of Stock Market Performance during election years

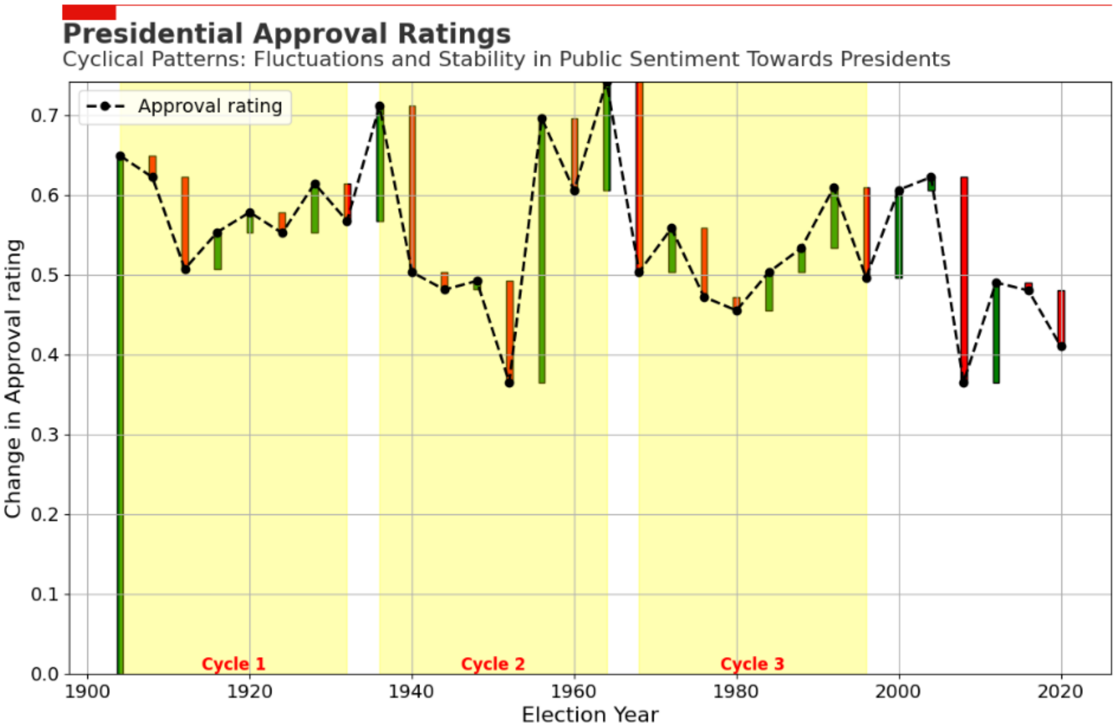


Figure A.5: Waterfall chart of Presidential Approval Ratings during election years

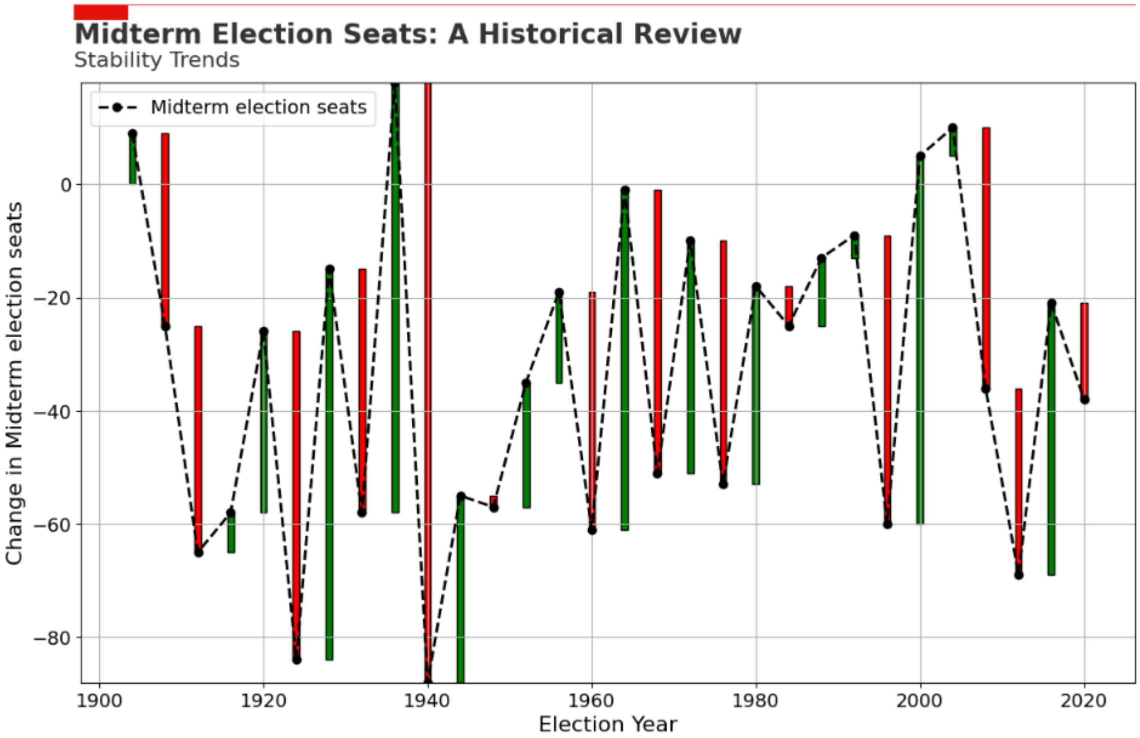


Figure A.6: Waterfall Chart of the Change in Incumbent Party’s Seats during Midterm Elections

Appendix B

Election Year	Incumbent party	Popular vote percentage	Inflation rate	GDP growth rate	Exchange Rate	Unemployment rate	stock market performance
1904	Republican	0,564	1,1	-3,11%	4,87	5,400	0,247
1908	Republican	0,516	-2,1	-9,95%	4,87	8,000	0,385
1912	Republican	0,232	2,1	3,07%	4,87	4,600	0,023
1916	Democratic	0,492	7,9	12,21%	4,77	5,100	-0,034
1920	Democratic	0,341	15,6	-2,26%	3,66	5,200	-0,190
1924	Republican	0,540	0,0	1,11%	4,42	5,000	0,240
1928	Republican	0,582	-1,7	-0,10%	4,87	4,200	0,376
1932	Republican	0,396	-9,9	-15,11%	3,51	23,600	-0,152
1936	Democratic	0,608	1,5	9,81%	4,97	16,900	0,279
1940	Democratic	0,547	0,7	8,11%	3,83	14,600	-0,153
1944	Democratic	0,534	1,7	6,56%	4,04	1,200	0,138
1948	Democratic	0,495	8,1	3,57%	4,03	3,750	-0,007
1952	Democratic	0,444	1,9	1,97%	2,79	3,025	0,118
1956	Republican	0,574	1,5	0,16%	2,80	4,125	0,026
1960	Republican	0,495	1,7	0,88%	2,81	5,542	-0,030
1964	Democratic	0,611	1,3	4,33%	2,79	5,158	0,130
1968	Democratic	0,427	4,2	3,72%	2,39	3,558	0,077
1972	Republican	0,607	3,2	4,18%	2,50	5,600	0,156
1976	Republican	0,480	5,8	4,25%	1,80	7,700	0,192
1980	Democratic	0,410	13,5	-1,13%	2,33	7,175	0,258
1984	Republican	0,588	4,3	6,36%	1,34	7,508	0,014
1988	Republican	0,534	4,1	3,27%	1,78	5,492	0,124
1992	Republican	0,374	3,0	2,13%	1,77	7,492	0,045
1996	Democratic	0,492	3,0	2,60%	1,56	5,408	0,203
2000	Democratic	0,484	3,4	2,94%	1,52	3,967	-0,101
2004	Republican	0,507	2,7	2,83%	1,83	5,542	0,090
2008	Republican	0,457	3,8	-1,23%	1,85	5,800	-0,385
2012	Democratic	0,511	2,1	1,45%	1,59	8,075	0,134
2016	Democratic	0,482	1,3	0,81%	1,35	4,875	0,095
2020	Republican	0,469	1,2	-2,59%	1,28	8,092	0,163

Table B.1: Economic data

Election Year	Incumbent party	Electoral percentage	Campaign Spending Index	Approval rating	Midterm election seats	Geopolitical rating	Running for reelection	Period in power
1904	Republican	70,6	1	0,649	9	0	1	1
1908	Republican	66,5	1	0,622	-25	2	0	1
1912	Republican	1,5	0	0,507	-65	1	1	1
1916	Democratic	52,2	1	0,553	-58	2	1	0
1920	Democratic	23,9	1	0,578	-26	3	0	1
1924	Republican	71,9	1	0,552	-84	2	0	0
1928	Republican	83,6	1	0,614	-15	1	0	1
1932	Republican	11,1	0	0,567	-58	1	1	1
1936	Democratic	98,5	1	0,711	18	0	1	0
1940	Democratic	84,6	1	0,503	-88	2	1	1
1944	Democratic	81,4	1	0,481	-55	1	1	1
1948	Democratic	57,1	1	0,492	-57	1	0	1
1952	Democratic	16,8	0	0,365	-35	3	0	1
1956	Republican	86,1	1	0,696	-19	1	1	0
1960	Republican	40,8	0	0,605	-61	4	0	1
1964	Democratic	90,3	1	0,742	-1	2	1	0
1968	Democratic	35,5	0	0,503	-51	2	0	1
1972	Republican	96,7	1	0,558	-10	2	1	0
1976	Republican	44,6	0	0,472	-53	2	1	1
1980	Democratic	9,1	0	0,455	-18	0	1	0
1984	Republican	97,6	1	0,503	-25	3	1	0
1988	Republican	79,2	1	0,533	-13	1	0	1
1992	Republican	31,2	1	0,609	-9	4	1	1
1996	Democratic	70,4	1	0,496	-60	2	1	0
2000	Democratic	49,4	0	0,606	5	2	0	1
2004	Republican	53,2	1	0,622	10	3	1	0
2008	Republican	32,2	0	0,365	-36	4	0	1
2012	Democratic	61,7	1	0,490	-69	3	1	0
2016	Democratic	43,1	1	0,480	-21	3	0	1
2020	Republican	43,1	0	0,410	-38	4	1	0

Table B.2: Non-economic data