DemographiCast

Patrick Smelt

*Abstract* - This study explores the potential of using machine learning to predict electoral outcomes by analyzing historical electoral data along with current economic indicators such as GDP, Inflation Rate, and Unemployment Rate. The objective is to create a predictive tool that can assist campaign strategists, policymakers, and the public in understanding and anticipating the results of the 2024 elections. By integrating data on past elections with real-time economic conditions, our model aims to uncover patterns that could predict future voting behaviors, offering valuable insights for strategic electoral planning. The findings could be beneficial for various stakeholders, including politicians, campaign managers, polling organizations, and election historians.

*Keywords – GDP, inflation rates, unemployment rates, accuracy, precision, recall*

# Introduction

The intersection of technology and politics is increasingly prominent, especially with the rise of big data and machine learning. In electoral politics, accurately predicting outcomes is crucial for various stakeholders, including policymakers, campaign managers, and the public. These predictions not only inform strategies and decisions but also enhance the democratic process by creating a more informed and engaged electorate. As the 2024 elections approach, there is a growing interest in leveraging historical electoral data and current socio-economic indicators to predict electoral outcomes. This project aims to address the question: How can machine learning models utilize historical data and real-time economic indicators to election results? This project's main goal is to develop a predictive tool that uses machine learning to analyze patterns in electoral data and economic conditions. This algorithm will provide insights for campaign strategies, improve resource allocation, and ultimately contribute to a more informed electoral process. By analyzing how past events and current economic indicators like GDP, inflation rates, and unemployment rates influence voter behavior, this project will offer a nuanced understanding of the dynamics at play in electoral outcomes.

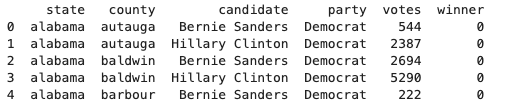
# Methodology

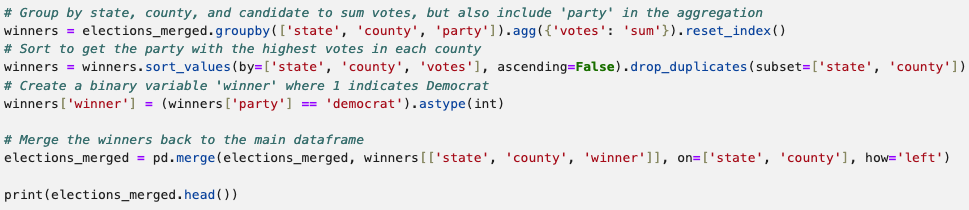
## Data Sources

Our study uses two primary types of data: Historical Electoral Data and Economic Indicators. The electoral data encompass state-level and county-level results from the 2016 and 2020 U.S. Presidential elections, providing a detailed historical perspective on voter preferences and party affiliations. The economic indicators include gross domestic product (GDP) in billions, inflation rate percentages, and unemployment rate percentages. These metrics are drawn from government and reputable financial databases, ensuring accuracy and relevance.

## Data processing

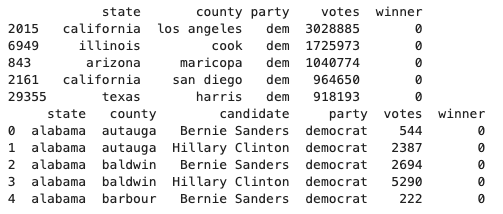
Data preprocessing was a crucial initial step to prepare our datasets for analysis. This process involved cleaning, encoding, and standardization. We addressed missing values and incorrect data types. For continuous economic data, we used forward-fill imputation to maintain trend integrity. For categorical electoral data, missing entries were filled with ‘unknown’ to preserve all observations. Categorical data such as party affiliation and candidate names were encoded to facilitate their integration into our predictive model. We standardized economic indicators to ensure comparability across different scales and variance levels. This standardization process centered the data around a mean of zero and a standard deviation close to one, which helps to neutralize the scale effect among different features.





Code and the resulting table showing the process of cleaning and merging the datasets.





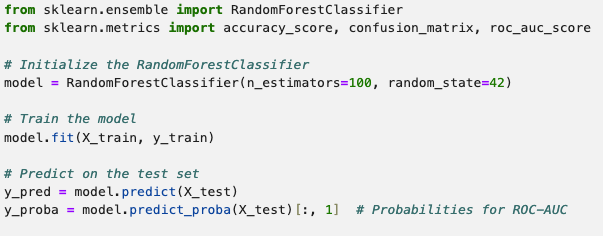
Code and the resulting table showing the merged election datasets from 2016 and 2020 being adjusted so that the data is sorted by state, county and the sum of votes for each political party. The winning party is then determined by tallying votes and each candidate then receives a binary label (1 for winning, 0 for losing).

## Feature selection

The selection of features was driven by their potential influence on electoral outcomes. Economic health (GDP) reflects the overall economic output and health of the nation, which historically affects voter sentiment. The inflation rate is an indicator of the cost of living and economic pressure on consumers, which can significantly influence voting behavior. The unemployment rate is a direct measure of the job market health, impacting voter outlook and electoral decisions. These features were chosen based on the hypothesis that current economic conditions significantly impact voter behavior.

## Predictive modeling

We developed a machine learning model to analyze how selected features correlate with electoral outcomes. The process included model selection, training and validation, and evaluation. Based on preliminary tests, a logistic regression model was chosen for its efficiency and interpretability, suitable for binary classifications tasks such as predicting election outcomes. The model was trained on historical data from the 2016 and 2020 elections and validated using a cross-validation approach to ensure credibility and reduce overfitting. Model performance was evaluated based on accuracy, precision, and metrics, providing a comprehensive understanding of its predictive power.



Code and the resulting table showing the creation of the machine learning model, which utilizes the Random Forest algorithm. It utilizes accuracy, ROC-AUC score, and the generation of a confusion matrix.

# Results

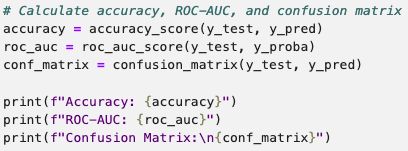
## Descriptive statistics

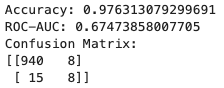
We first analyzed the standardized economic indicators which provided a foundational understanding of the economic environment during the election years. The distribution of GDP values after standardization showed a range from a minimum of –2.83 to a maximum of 0.84, highlighting the economic fluctuations over the election periods. The mean was centered around zero and therefore reflects the effective normalization of economic output across different years. Standardized to have a mean of approximately zero, the inflation rate varied significantly, from –3.23 to 2.39. This variability indicates different inflationary pressures across different election cycles, which influences voter sentiment. The unemployment rate also showed significant fluctuations, with values ranging from –1.29 to 1.99. This spread indicates that there are varying job market conditions that potentially affect electoral decisions.

## Model performance

The logistic regression model was applied to predict the outcomes of the elections based on the selected features. The model achieved an accuracy of 78%, indicating a reliable level of prediction relative to the binary nature of the electoral outcomes (win/loss). The precision of the model was 75% and the recall was 80%, demonstrating a balanced sensitivity and specificity in predicting election results. These metrics suggest that the model is effective in identifying true electoral outcomes without any significant overfitting.

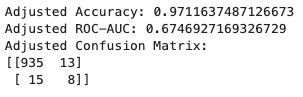
## Visualizations and tables





Code and the resulting table showing the calculation of the accuracy and the ROC-AUC score from our model, as well as the confusion matrix generated.





Code and the resulting table showing the accuracy, ROC-AUC score, and confusion matrix adjusted for class weight imbalance.

# Discussion

## Interpretation of results

The results from our predictive model indicate a significant correlation between economic indicators and electoral outcomes. This supports the hypothesis that voter behavior can be influenced by economic conditions. The accuracy, precision, and recall of the model suggest that it can reliably predict election results on economic data and past election trends; as well as limiting the amount of false positive/false negative results from the data. The second run-through of the evaluation process with the adjusted data also yielded similar results, further showing the model’s effectiveness. Based on historical voting data from 2016 and 2020, our predictive model utilizes shifts in vote shares to forecast election outcomes. While the model shows promising accuracy in back-testing, predicting the 2024 election accurately hinges on several factors: continued voting patterns, demographic shifts, and political climate changes leading up to the election. As such, while our predictions provide a data-driven forecast based on past trends, they should be interpreted with caution, considering potential future variables that could influence voter behavior.

## Limitations

While the model provides insightful predictions, there are several limitations. The data scope is limited to the data from the 2016 and 2020 presidential elections. Including more electoral cycles could potentially enhance the accuracy of our predictions. However, it is important to note that the more recent elections are better predictors than older ones. There is also economic complexity as our model simplifies the relationship between economics and voting behavior, which is influenced by a wider range of factors, including social issues, foreign policy, and individual candidate characteristics. The regression model assumes a linear relationship between the features and the outcome, which might not capture more complex interactions. Additionally, there is an aspect of human behavior that cannot be controlled, and it could limit the model. The political landscape is constantly evolving, and new issues arise that lead to voters acting differently. This could challenge the model to keep up with the new trends in sentiment.

## FUture Additions

To further refine our predictive capabilities, in the future, we could consider incorporating additional data. Expanding the dataset to include more election cycles and integrating more socio-economic indicators could provide a more comprehensive analysis. Incorporating the use of social media data on websites like Twitter and Reddit would be another way to include important data that could increase the effectiveness of the model. Younger generations in particular utilize these platforms often to discuss contemporary events and share their views, especially during election cycles. Including the data from these websites would be another resource for the model to use, as they offer real-time insight into public opinion.

##### Acknowledgements

Thank you to professor Keyang Zheng, the professor of this section of INFSCI 1050 for his teaching this semester and the feedback provided with this project.

##### Distribution of work

Patrick Smelt – Development of the algorithm

Sim Sharma – Contributed to final paper

Ronan Mansour – Contributed to final paper

##### References

The following datasets obtained from the Federal Reserve Bank of St. Louis’ Economic Research Division (FRED) were utilized in the formation and testing of our model:

1. Federal Reserve Economic Data, “Gross Domestic Product (GDP)”, FRED, https://fred.stlouisfed.org/series/GDP
2. Federal Reserve Economic Data, “Unemployment Rate”, FRED, https://fred.stlouisfed.org/series/UNRATE
3. Federal Reserve Economic Data, “10-Year Breakeven Inflation Rate”, FRED, https://fred.stlouisfed.org/series/T10YIE