Unveiling Voter Trends: Data Analysis and Predictive Modeling for the 2024 US Election

Aneesh Ojha
Electrical and Computer Engineering
University of California, San Diego
anojha@ucsd.edu

Margi Pandya
Electrical and Computer Engineering
University of California, San Diego
mapandya@ucsd.edu

***Abstract*—The 2024 US Presidential Election features a complicated mix of voter behavior, political polarization, and changing demographic trends. This project uses data analysis and machine learning techniques to study underlying voter behavior trends as well as create predictive models that shed light on prospective election outcomes. We created a comprehensive forecasting system by incorporating several data sources, such as historical vote records, demographic data, social media mood, and current polling findings. The study's goal is to improve our understanding of voter behavior through exploratory data analysis (EDA), statistical methodologies, and data visualization. We aim to discover the key elements influencing vote preferences by carefully studying links within election data.**

***Index Terms*—Exploratory data analysis (EDA), Machine learning, Data visualization, Demographic data and trends**

I. Introduction

The landscape of American politics has changed a lot in recent years, with increased complexity and quickly changing demographic patterns. Various factors influence the outcome of U.S. elections, shaping both voter behavior and electoral results. Key factors include voter turnout, party dynamics, and the political environment, which are often influenced by economic conditions and national events. Media coverage and campaign strategies also play critical roles in framing issues and mobilizing voters. Additionally, the growing diversity of the American population, along with rapidly evolving socioeconomic conditions, requires more sophisticated methods for analyzing voter behavior.

This project focuses on understanding the factors that influenced the outcomes of both the 2020 and 2024 presidential elections, with a particular emphasis on exploring how demographic, political, and social trends shaped voter preferences. The study highlights the victory of the Republican Party in 2024 and seeks to explore the underlying forces that drove this result, following the contentious 2020 election.

Data exploration plays a central role in this project, as it allows for the identification of key patterns, trends, and anomalies in the election data. By analyzing election results at the state level, the study provides a granular view of voting behavior across different regions of the United States. Additionally, U.S. Census data is incorporated to enrich the analysis by offering insights into the demographic, gender and race-based support for each political party. The project employs a variety of techniques for data exploration, starting with the cleaning and manipulation of raw data using Python libraries such as Pandas and NumPy. Univariate analysis is used to examine individual variables in the census data and the political data, while multivariate analysis uses statistics-based methods to understand the relationships among the dependent and independent variables.

Ultimately, this project provides a comprehensive analysis of the 2020 and 2024 U.S. presidential elections, offering valuable insights into the evolving political landscape of the United States. Through data exploration and visualization, the study sheds light on the key factors that shaped voter preferences, providing a clearer understanding of how the U.S. electorate has changed over time. Additionally, machine learning algorithms such as linear regression, K-Nearest Neighbors (KNN) regression, and Random Forest regression are employed to make predictions for swing states, analyzing how demographic factors within these states may influence election outcomes. This approach enables a deeper understanding of how state-specific characteristics impact the results and offers a predictive model for future elections. We have uploaded the code files and datasets used for this project in this link: GitHub Repository.

II. Methodology

A. Data Exploration

The data used in the analysis contains demographic and voting information from the 2020 and 2024 U.S. elections dataset [1], [3] and the US Census dataset [2]. Key demographic variables include raw population counts for groups such as Men, Women, Hispanic, Black, and Asian populations, along with the total population for each county or region. These raw counts are normalized to calculate demographic proportions by dividing them by the total population. Normalization ensures that the analysis focuses on relative demographic influences rather than absolute population sizes.

Voting data [1], [3] are represented through two key variables: the percentage of votes estimated for the Republican Party and the Democratic Party. Preprocessing steps include handling missing values by replacing them with zeroes and creating proportion-based metrics. Statistical analyses, such as logistic regression and correlation, compare Republican party voting percentages between counties with ethnic composition and gender, where a threshold is used to classify these groups.

Visualizations are used to examine relationships between demographic proportions and Republican voting percentages and bar plots are used to depict state-level averages of demographic proportions and voting trends. This dataset, with its rich combination of demographic and voting information, serves as the foundation for exploring the relationships between population composition and election outcomes.

1) Statistical Test: The logistic regression model is widely used in data exploration because it is an effective tool for analyzing relationships between a binary outcome variable and one or more predictor variables. Its primary purpose is to model the probability that a certain event occurs, which makes it suitable for classification problems.

The logistic regression model is defined by the following equation:

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

where $P(1 \mid X)$ (i.e. $P(y = 1 \mid X)$) represents the probability that the dependent variable y equals 1 given the predictors X, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the predictors $X_1, X_2, \ldots, X_n$, which are the independent variables.

The model is estimated by maximizing the log-likelihood function, which is given by

$$L(\beta_0, \ldots, \beta_n) = \prod_{i=1}^{N} \left(P(1 \mid X_i)^{y_i} \cdot (1 - P(1 \mid X_i))^{1-y_i} \right)$$

$$\log L_1 = \sum_{i=1}^{N} y_i \log P(y_i = 1 \mid X_i) \tag{1}$$

$$\log L_2 = \sum_{i=1}^{N} (1 - y_i) \log(1 - P(y_i = 1 \mid X_i)) \tag{2}$$

Substitute equations 1 and 2 into the equation 3

$$\log L(\beta_0, \beta_1, \ldots, \beta_n) = \log L_1 + \log L_2 \tag{3}$$

where N is the number of observations and y_i represents the observed binary outcome for observation i.

This function is maximized to find the optimal values for the coefficients $\beta_0, \beta_1, \ldots, \beta_n$.

Once the coefficients are estimated, the model can be evaluated using various metrics, including:

- **Pseudo R-squared [5]**: A measure of goodness-of-fit for logistic regression, similar to R-squared in linear regression.
- **Likelihood Ratio Test [6]**: Compares the fit of two models, often used to compare nested models.

The odds ratio for each predictor can be computed as the exponential of the coefficient:

$$\text{Odds Ratio} = e^{\beta}$$

This indicates how the odds of the event (e.g., $y = 1$) change with a one-unit increase in the predictor.

B. Regression Analysis

Our electoral prediction study uses machine learning algorithms to forecast voting probabilities for both the Republican and Democratic parties in key swing states. We used regression models based on polling estimates, candidate momentum, electoral leads, and historical voting patterns by randomly partitioning our full dataset into training and testing groups. This method enables us to acquire insight into voter preferences.

We use actual vote-share data from past elections up to 2020 as the foundation for training our models. By examining statistical relationships between various variables, we tried to project potential election outcomes.

1) Models: In this project, we used three models: Linear Regression, KNN regression, and Random Forest.

a) Linear Regression: Linear regression works by finding the best straight line that predicts an outcome based on one or more input variables. The basic idea is to make predictions by using a formula that relates the input (like polling data) to the output (like voting percentage). The formula for a simple linear regression model is:

$$y = (\beta_0 + \beta_1 x + \epsilon) \tag{4}$$

Here, y is the predicted value, x is the input (like polling data), β_0 is the starting point of the line and β_1 is the slope, which tells us how much y changes when x changes. The goal is to find the values of β_0 and β_1 that make the predictions as close as possible to the actual outcomes. This is done by minimizing the difference between predicted and actual values, which is calculated using a formula that adds up these differences squared. The loss function is given by:

$$\text{Loss} = \sum (y_{\text{actual}} - y_{\text{predicted}})^2 \tag{5}$$

The algorithm tries to make this loss as small as possible by adjusting β_0 and β_1 until the line fits the data points as well as it can. The algorithm tries to make this loss as small as possible by adjusting β_0 and β_1 until the line fits the data points as well as it can.

b) KNN(K-Nearest Neighbors) Regression: K-Nearest Neighbors (KNN) regression works by predicting a value based on the average of the nearest data points to the one we want to predict. Instead of finding a line, like in linear regression, KNN looks at the kk closest data points and uses them to make a prediction. The number kk is the number of neighbors we consider. The formula for making a prediction is:

$$y_{\text{predicted}} = \frac{1}{k} \sum i = 1^k y_i \tag{6}$$

Where y_i are the actual values (such as voting percentages) of the k closest points, and $y_{\text{predicted}}$ is the average of these values. The algorithm measures how close the data points are, often using the distance between them (such as Euclidean distance). Once it finds the nearest points, it takes the average

of these points to predict the value for the new data point.

c) Random Forest: Random Forest regression works by creating many decision trees and then averaging their predictions to make a final result. Each decision tree looks at the data, splits it based on certain features, and makes a prediction. The Random Forest algorithm combines the predictions from all the trees to get a more accurate result. The formula for the final prediction model is given by the equation:

$$y_{\text{predicted}} = \frac{1}{T} \sum_{t=1}^{T} y_t \tag{7}$$

Where y_t is the prediction from each individual tree, and T is the total number of trees. Each tree in the forest makes its own prediction, and the final result is the average of all those predictions. This method helps reduce errors and improve accuracy by using the wisdom of many trees instead of relying on just one.

III. Implementation Details

A. Dataset

The "US Election 2020" [1] dataset on Kaggle provides comprehensive election results for the 2020 Presidential Election, including vote counts for major political parties (Democrats and Republicans), voter turnout, and demographic information such as age, gender, and ethnicity across various states and counties. In conjunction, the "US Census Demographic Data" [2] offers detailed socio-economic insights, including data on age, race, gender, household composition, income, and employment across U.S. states, counties, and districts. Combined, these datasets enable in-depth analysis of voting patterns, demographic influences, and socio-economic trends, providing a richer context for political research and understanding election outcomes.

Our election project uses detailed polling data for the 2024 US presidential election from the website FiveThirtyEight [3], as well as actual state-by-state voting results provided by the Federal Election Commission (FEC). We focus specifically on key swing states, selecting data from Pennsylvania, Wisconsin, Michigan, Georgia, North Carolina, Arizona, and Nevada. Because the accuracy and availability of data have improved significantly since 2000, our analysis focuses on the six most recent election cycles. We also considered party dominance as a feature in our model, which helps improve predictions by recognizing historical voting trends and public sentiment towards the party currently in power.

IV. Results

A. Exploratory Dataset Analysis Comparison

1) Vote Margin: The vote margin in an election refers to the difference in the number of votes between two candidates or political parties. In the context of the U.S. 2020 Presidential Election, as shown in Figure 1, the District of Columbia (DC) showed a clear preference for the Democratic Party, with a significant vote margin favoring the Democrats. On the other hand, Wyoming, a traditionally Republican-leaning state, recorded the highest vote margin for the Republican Party

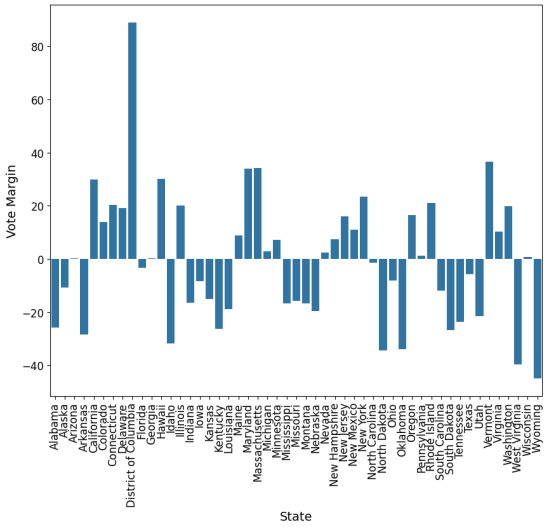

Fig. 1. shows the vote margin plot for the 2020 Presidential election.

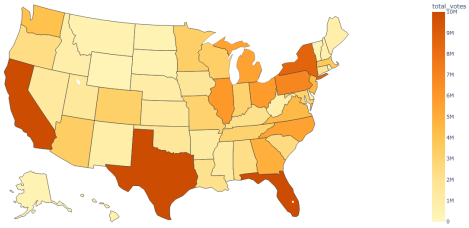

Fig. 2. shows the vote turnout plot for the 2020 Presidential election code using the Plotly library to create a choropleth map of the USA, visualizing the total votes in each state with color intensity.

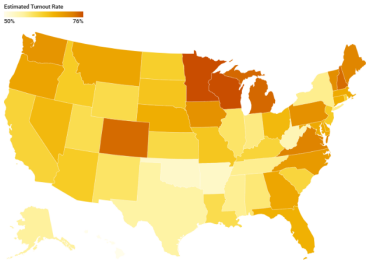

Fig. 3. shows the vote turnout plot for the 2024 Presidential election code estimated by US News [4]

2) Vote Turnout: 2024 has seen a slight decrease in voter turnout from 2020, mainly due to the pandemic-induced surge in mail-in ballots in 2020, but still higher than pre-2020 elections. An estimated 64% by [4] of the voting-eligible population voted, with about 155 million ballots cast. Approximately 89 million Americans (or 36% of eligible voters) did not vote. One of the first observations was the variation in voter turnout across different states. As the figure 1 amd 2 shows the states with larger urban populations, such as California, New York, and Texas, consistently had higher voter turnout.

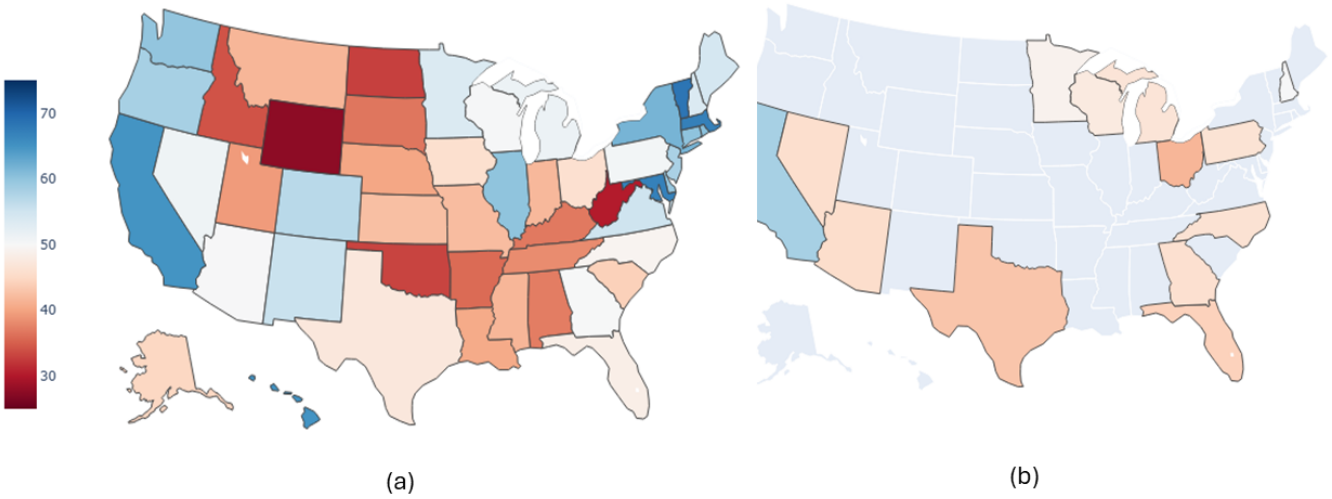

Fig. 4. (a) shows the choropleth map of the United States, visualizing the percentage of the population in each state that voted for the Democratic Party in the 2020 presidential election. (b) illustrates the choropleth map of the United States based on the estimated dataset, predicting the possible swing states. The results of (b) highlight only the swing states and major states, due to the use of exit poll dataset.

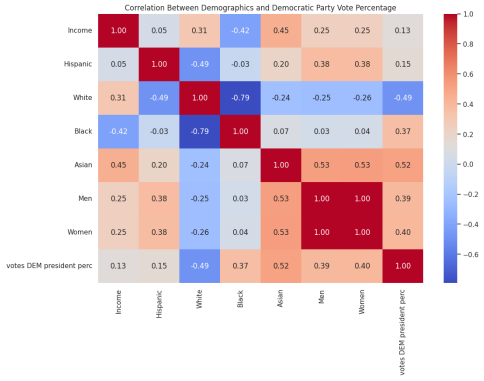

Fig. 5. The heatmap highlights key correlations between demographics and the Democratic vote percentage in the swing states in the 2020 elections.

Additionally, there was a noticeable increase in total voter participation in 2024 compared to 2020, reflecting growing political engagement.

States like Minnesota and Wisconsin had record-high turnout in 2020, partly due to the expansion of voting options. In 2024, while states like Minnesota and Wisconsin still had high turnout (above 75%), other swing states like Arizona, Georgia, and Michigan saw relatively lower turnout compared to 2020 observed while data exploration, when the stakes were higher and mobilization efforts were more intense.

a) Swing States: States like Michigan, Pennsylvania, Wisconsin, and Arizona, which were crucial in determining the 2020 election, remained competitive in 2024. Apart from them Nevada and North Carolina were added as shown in Figure 4. However, it was estimated a clear shift in voter preferences in these states, with a notable increase in Republican support.

B. 2020 Presidential Election Data Exploration

1) Demographic Trends: The analysis of U.S. Census data revealed how demographic factors influenced voting patterns, especially in terms of gender and ethnicity. The exploration highlighted the growing importance of minority groups and women in shaping election outcomes.

a) Gender: The gender divide in voting patterns became more pronounced. Women, particularly those in suburban areas, leaned more towards the Democratic Party, while men showed stronger support for the Republicans.

The results of the coefficients of the logistic regression model for gender features are as shown in Table I. The logistic regression model, with a pseudo-R-squared of 0.1628, explains approximately 16% of the variation in the percentage of votes for the Democratic Party. The model is statistically significant (LLR p-value = 3.576×10^{-202}) and performs better than the null model. The coefficient for Men_perc is 5.8080, indicating that an increase in the percentage of men is strongly associated with a higher percentage of votes for the Democratic Party. Conversely, Women_perc has a coefficient of -3.8926, showing that a higher percentage of women is associated with a decrease in Democratic votes. Both predictors are highly significant (p-value = 0.000), and the intercept (-0.3419) represents the baseline when all predictors are zero. These results highlight the significant impact of gender proportions on voting for the Democratic Party.

b) Ethnic Composition: Minority groups, especially Black, Hispanic, and Asian American voters, showed consistent support for the Democratic Party in both elections, though the Republican Party saw increased engagement from Hispanic voters in 2024, especially in states like Texas and Florida.

The logistic regression results in Table VII suggest a slight positive relationship between the Hispanic population proportion (Hispanic_perc) and the likelihood of voting Republican,

TABLE I
LOGISTIC REGRESSION RESULTS FOR GENDER FEATURE IN 2020 ELECTIONS

Variable	Coefficient	Std. Error	z-Value	P-Value	[0.025, 0.975]
const	-0.3419	0.051	-6.715	0.000	[-0.442, -0.242]
Men_perc	5.8080	0.762	7.624	0.000	[4.315, 7.301]
Women_perc	-3.8926	0.769	-5.059	0.000	[-5.401, -2.385]

TABLE II
LOGISTIC REGRESSION RESULTS FOR ETHNIC COMPOSITION FEATURE IN 2020 ELECTIONS

Variable	Coefficient	Std. Error	P-Value	[0.025, 0.975]
const	0.6391	0.037	0.0000	[0.566, 0.713]
Hispanic_perc	0.0018	0.000	0.0000	[0.001, 0.003]
Black_perc	-0.0009	0.000	0.0000	[-0.001, -0.000]
Asian_perc	0.0355	0.005	0.0000	[0.025, 0.046]

TABLE III
LOGISTIC REGRESSION RESULTS FOR ETHNIC COMPOSITION FEATURE IN 2024 ELECTIONS

Variable	Coefficient	Standard Error	Z-Statistic	P-Value
Intercept (const)	5.9211	5.656	1.047	0.295
Hispanic_perc	2.2778	2.167	1.051	0.293
Black_perc	2.2584	1.734	1.303	0.193
Asian_perc	-49.2685	39.316	-1.253	0.210

though the effect size is small. The coefficient of 0.0018 indicates that as the Hispanic population increases, the probability of voting Republican slightly increases, but the effect is modest.

On the other hand, counties with higher Black population proportions (Black_perc) are less likely to vote Republican, with a negative coefficient of -0.0009. This aligns with historical voting patterns, where Black communities tend to lean Democratic. Conversely, counties with higher Asian population proportions (Asian_percent) are more likely to vote Republican, with a coefficient of 0.0355. This trend, while positive, maybe region- or state-specific, highlighting potential demographic differences in voting behavior.

c) Swing States: One of the most critical findings from the data exploration was the shift in support within key swing states. States like Michigan, Pennsylvania, Wisconsin, and Arizona, which were crucial in determining the 2020 election. The demographic trends in swing states were similar to overall trends. Asian and Black populations show positive correlations with Democratic voting, with Asian demographics particularly strong (0.52). Income is positively correlated with Asians but negatively with Black demographics. White demographics are negatively correlated with the Democratic vote (-0.49), while the Hispanic population is negatively correlated with White and positively with Women demographics. Men and Women are perfectly correlated with each other, however, there are no significant gender differences in voting behavior in swing states. These findings emphasize the demographic factors that influence Democratic voting patterns.

C. 2024 Presidential Election Data Exploration

1) Demographic Trends:

a) Ethnic Composition: The logistic regression model shows a moderate fit with a pseudo-R-squared of 0.6222 as illustrated in Table III, indicating that the model explains about 62% of the variation in the dependent variable. However, none of the predictor variables—Hispanic, Black, or Asian population percentages—are statistically significant, with p-values greater than 0.05. The coefficients suggest positive relationships for Hispanic and Black populations and a strong negative relationship for the Asian population, but the large standard errors and non-significance indicate uncertainty. Additionally, the model faces issues of quasi-separation, where 23% of observations can be perfectly predicted, leading to potential instability in parameter estimates.

There is no complete dataset available for the 2024 US presidential election; hence, the data file being used is estimated by a model from a news channel [4]. Therefore, the results generated by the logistic regression test exhibit infinite variance. The model becomes too sensitive to small fluctuations in the training data, leading to unstable predictions. The test results for gender features are not ignored.

TABLE IV
RANDOMFOREST REGRESSION VOTE SHARE

State	DEM	REP
Arizona	48.919269	49.013539
Georgia	49.300137	49.432486
Michigan	53.993855	44.701297
Nevada	46.084982	49.011648
North Carolina	49.404060	48.989135
Pennsylvania	51.140649	48.393981
Wisconsin	51.515807	45.620168

D. Prediction

The KNN Regression, Linear Regression, and Random Forest models showed varying accuracy in predicting the 2024 election vote share in swing states, consistently under-

TABLE V
LINEAR REGRESSION VOTE SHARE

State	DEM	REP
Arizona	49.150266	49.190028
Georgia	49.283083	49.354297
Michigan	49.803704	48.700209
Nevada	48.791889	48.583514
North Carolina	49.281091	49.640685
Pennsylvania	49.466730	48.998972
Wisconsin	50.005884	48.327966

TABLE VI
KNN VOTE SHARE

State	DEM	REP
Arizona	48.979024	48.939237
Georgia	49.322030	49.354754
Michigan	53.052056	46.174231
Nevada	48.953806	47.831246
North Carolina	49.540650	49.546872
Pennsylvania	50.952342	48.280634
Wisconsin	52.006918	48.023556

TABLE VII
2024 ACTUAL VOTE SHARE

State	DEM	REP
Arizona	47.900	52.100
Georgia	49.300	50.700
Michigan	50.300	49.700
Nevada	49.300	50.700
North Carolina	49.100	50.900
Pennsylvania	49.600	50.400
Wisconsin	50.400	49.600

estimating the Republican vote share in key states like Arizona, Michigan, and Nevada, where actual Republican support exceeded predictions. For example, all models projected a Republican vote share of around 48.9%-49.0% in Arizona compared to the actual 52.1%, and similar underpredictions occurred in Nevada. Democratic vote share predictions were generally more accurate, though KNN and Random Forest often overestimated in states like Michigan and Wisconsin, while Linear Regression was closer to the actual results. KNN performed well in Georgia and Pennsylvania, Random Forest was more aligned with Republican outcomes in Nevada and North Carolina, and Linear Regression showed balanced predictions in Michigan and Wisconsin. Overall, the models performed reasonably but require refinement, particularly in capturing shifts in Republican vote shares, to improve future election forecasts.

V. CONCLUSION

The exploratory data analysis (EDA) of the 2020 U.S. Presidential Election revealed key demographic trends influencing voter behavior, such as men favoring Democrats and suburban women leaning Republican. Ethnic minorities, including Black, Hispanic, and Asian voters, showed strong Democratic support, though Republican engagement with Hispanic voters increased in 2024. Voter turnout varied by region, with urban areas showing higher participation and swing states like Michigan and Arizona experiencing notable shifts in preferences, underscoring demographic factors' impact. However, the 2024 elections exposed significant challenges in predictive modeling, as methods like KNN Regression, Linear Regression, and Random Forest struggled to account for rapid shifts in voter sentiment and polling inaccuracies. Polling flaws, such as sampling biases and non-response concerns, were compounded by varied techniques and data presentation in the media, reducing forecasting dependability. These issues highlighted the fundamental difficulty of predicting elections in a volatile political setting.

REFERENCES

[1] "2020 US Election Dataset," Kaggle, Accessed: Dec. 14, 2024. [Online]. Available: https://www.kaggle.com/datasets/unanimad/us-election-2020

[2] "US Census Demographic Data," Kaggle, Accessed: Dec. 14, 2024. [Online]. Available: https://www.kaggle.com/datasets/muonneutrino/us-census-demographic-data/data

[3] "Polls Data," GitHub, Accessed: Dec. 14, 2024. [Online]. Available: https://github.com/fivethirtyeight/data/blob/master/polls/README.md

[4] A. L. Benson, "How Many People Didn't Vote in the 2024 Election," US News, Nov. 15, 2024. [Online]. Available: https://www.usnews.com/news/national-news/articles/2024-11-15/how-many-people-didnt-vote-in-the-2024-election.

[5] Hu, Bo Shao, Jun Palta, Mari. "PSEUDO-R 2 in logistic regression model". Statistica Sinica. 16. 847-860, 2006

[6] Neyman, J., Pearson, E. S. . "On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character", 221(594), 289-337, 1928

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] FiveThirtyEight. "Who Is Favored To Win The 2024 Presidential Election" Nov. 5, 2024. [Online]. Available: https://projects.fivethirtyeight.com/2024-election-forecast/