Logistic Regression Analysis of Individual-Level Vote Choice in Recent U.S. Presidential Elections

Talia Fabregas

August 25, 2025

Abstract

This project examines individual-level vote choice recent American presidential elections, employing frequentist, machine learning, and Bayesian logistic regression methods. Model selection via likelihood ratio testing indicates that interactions between race and key demographic factors (gender, education, region, and the urban-rural divide), along with approval of then-incumbent President Joe Biden and perceptions of economic conditions significantly improve logistic model fit on the 2024 Cooperative Election Study (CES) dataset. An L2-regularized logistic regression model trained and optimized via minibatch gradient descent on the 2024 CES dataset achieved 92.8% accuracy on unseen test data. To further assess generalizability, the same model was re-trained on the 2020 Cooperative Election Study (CES) and 2016 Cooperative Congressional Election Study (CCES) datasets using analogous predictors and the same hyperparameters, and achieved 96.52% and 93.53% accuracy, respectively, on unseen test data. Findings suggest that the interaction effects of race with demographic and geographic factors, as well as economy-related perceptions and approval of the incumbent president, significantly influenced vote choice in recent U.S. presidential elections and that a voter-level logistic modeling approach that incorporates micro- and macro-level predictors is a step towards an aggregate election forecast.

Code and data can be found using this link: https://github.com/taliafabs/LogisticVoteChoice.git

1 Introduction

On November 5, 2024, Donald Trump was elected the 47th President of the United States, defeating then-Vice President Kamala Harris 312 to 226 in the electoral college, and 49.9% to 48.4% in the popular vote [1]. President Trump won all seven key battleground states – Michigan, Wisconsin, Pennsylvania, North Carolina, Georgia, Arizona, and Nevada – by close margins. Trump is a twice-impeached, 34- time convicted felon who became the first Republican nominee to win the popular vote since former President George W. Bush in 2004. His second victory followed an election cycle like no other. On July 21, 2024, then-President Joe Biden made the unprecedented decision to end his re-election campaign, following a disastrous debate performance, low approval ratings, and poor polling numbers [2]. Harris Harris quickly became the Democratic nominee with Biden's endorsement and ran an unprecedented 107-day campaign, the shortest in modern American history [3].Top election issues included inflation, the economy, the cost of living, medicare and medicaid, and reproductive rights [4, 5].

Since the November election, the following question has loomed: why did Americans re-elect Trump, despite all his baggage? This paper will investigate the question of why Americans voted the way that they did on November 5, 2024. This project aims build on the works of Kuriwaki et al. [6], Algara et al. [7], and Camatarri [8] by incorporating the interactions between race and education, race and region, and race and the urban-rural divide as well as voters' perception of the economy and incumbent presidential approval in a logistic regression vote choice model that focuses on individual voters. The resulting 2024 vote choice model will then applied to 2020 data to evaluate its ability to generalize. The estimands, whose exact value can never be known, are the true effects of the selected micro- and macro- level factors on the 2024 presidential vote choice. Third-party presidential candidates received only 1.85% of the national popular vote in 2024, therefore the outcome of interest – 2024 presidential vote choice – is treated as binary [9]. This project contributes to individual-level logistic vote choice modeling research by treating it as a binary classification task and incorporating both macro-level variables such as incumbent presidential approval and economic conditions and micro-level sociodemographic variables [8]. Instead of aiming to forecast an election outcome, this project aims to build and train a logistic regression model that can accurately classify likely Democratic and Republican presidential voters using the 2024 Cooperative Election Study (CES) dataset and achieve high predictive accuracy on unseen data. The Python programming language and the `numpy`, `pandas`, `scikit-learn`, `matplotlib`, `seaborn`, `statsmodels`, `TensorFlow`, and `keras` packages were used to produce the visualizations and train, fit, and evaluate the logistic models presented in this paper [10].

The remainder of this paper is structured as follows. Section 2 contains a brief discussion about the research presented by Kuriwaki et al. [6] on race and voting behavior, Algara et al. [7] on presidential approval and voting behavior, and Camatarri [8] on the use of logistic regression to model individual-level voting behavior that this project aims to build on. Section 3 contains an overview of the 2024 Cooperative Election Study (CES) Common Content Dataset that was used to train and evaluate logistic vote choice models and exploratory data analysis findings. Section 4 covers my modeling strategy and frequentist model selection process. Section 5 presents likelihood ratio test results and how well the selected model performed on unseen data.

2 Literature Review

Existing literature shows that presidential popularity and economic conditions are significant predictors of the outcomes of presidential elections [7, 8, 12]. Abramowitz [12] found that the more popular an incumbent president is and the better the state of the economy, the greater the vote share for the incumbent president's party should be [12]. More recently, Camatarri [8] built a logistic model to estimate support for Republican presidential nominees using pre-electoral waves of the ANES Time Series Study from 2012, 2016, and 2020 and found that economic disapproval was one of the strongest predictors for support of the Republican nominee when the incumbent president was a Democrat.

Race is a predictor of vote choice, but its significance is declining and should be considered in conjunction with other demographic and geographic variables. In "The Geography of Racially Polarized Voting: Calibrating Surveys at the District Level," Kuriwaki et al. [6] performed multilevel regression with post-stratification (MRP) using 2016 Cooperative Congressional Election Study (CCES), 2020 Congressional Election Study (CES) and U.S. Census data, hierarchal modeling with calibration weights to match their estimates to aggregated vote totals, and compared their estimates with those obtained using different techniques. They found that national-level racial group differences explain 60% of differences in district-level voting behavior and that race explains voting behavior more than geography, but its influence may be declining relative to education and the urban-rural divide [6].

Harris' chances in 2024 were hindered by Biden's low approval rating and poor public perceptions of economic conditions. In "Partisan Collective Accountability during the 2024 US Presidential and Congressional Elections," Algara et al. built a model to forecast four election outcomes of interest: presidential popular vote, presidential electoral college votes, number of U.S. Senate seats won by the incumbent president's party, and number of U.S. house seats won by the incumbent president's party using presidential approval and incumbent party brand as its two main predictors [7]. They found that presidential approval is the only key covariate that predicts the popular vote percentage and number of electoral votes won by the nominee from the incumbent president's party [7]. This suggests that both these factors influence individual-level voting behavior. Algara et al. validated the accuracy of their forecasting model by applying it to past U.S. elections; it would have correctly predicted the electoral college winner in 15 of 21 presidential elections since 1940 [7]. Their model correctly forecasted that Harris would narrowly lose the popular vote and the electoral college to Trump due to Biden's low presidential approval rating [7].

While macro-level factors such as incumbent presidential approval and economic conditions have significant influence on overall election outcomes, understanding individual-level voting behavior requires models that incorporate individual voter characteristics. In "Predicting Popular-vote Shares in US Presidential Elections: A Model-based Strategy Relying on Anes Data," Camatarri proposes the use of standard and Bayesian logistic regression models that consider voter-level data – such as political positions, ideology, race, gender, and other socioeconomic variables – for election predictions that are based on individual-level vote choice [2, 8]. Election forecasting models that only use macro-level variables – such as economic conditions and incumbent presidential approval – aim to predict a series of many micro-level outcomes: individual voters' decisions [8]. Camatarri's models were tested on the 2012, 2016, and

2020 pre-electoral waves of the ANES Time Series study and 10 of the 12 models had a forecasting error below 3% [8]. These results suggest that regression-based individual voter-level election models have the potential to produce accurate popular vote forecasts and compliment macro-level models [8].

Accurately classifying likely Democratic and Republican voters is a precursor to an individual-level logistic modeling approach that aims to forecast aggregate presidential election outcomes. Neither Kuriwaki et al. nor Algara et al. focus specifically on logistic vote choice modeling, but their findings highlight the influence of race on voting behavior and the predictive power of incumbent presidential approval in forecasting election outcomes [6, 7]. This project builds on Camatarri's [8] approach by including macro- and micro-level predictors in a logistic vote choice model trained on the 2024 CES dataset to estimate how these factors influenced the American electorate – comprised of individual voters – to send Trump back to the White House.

3 Data

3.1 Dataset overview

The 2024 Cooperative Election Study (CES) Common Content was the primary dataset used for this project [13]. The data was most recently downloaded via url from Harvard Dataverse in July 2025. The 2024 CES is an election-year wave of the CCES; it contains data from over 50,000 online interviews with American adults, conducted in the weeks leading up to and the weeks following the November 5th, 2024 presidential election [14]. The pre-election wave includes questions about general political attitudes, demographics, positions on current policies being debated in Congress, political information, and vote intentions; the post-election wave includes questions about how respondents voted in the recent election, and recent political and social engagement [13, 14]. Additional data preparation details are available in Appendix A.

The 2020 CES [15] and 2020 Cooperative Congressional Election Study [16] were used for the second iteration of this project. Both datasets are additional election-year waves of the study [14].

3.2 Exploratory data analysis

Exploratory data analysis findings suggest that education, race, geographic, incumbent presidential approval, and economy-related variables should be included in a logistic vote choice model. Harris voters are over-represented in the 2024 CES survey dataset. Visualizations are available in Appendix A.2. As shown in Figure 4 the 2024 CES survey dataset is Harris +12; this is not representative of the actual results, where Trump won the popular vote by 1.5 percentage points. However, the dataset includes post-stratification weights to account for the under-representation of voters from strata associated with support for Trump [13]. As shown in 2, higher education is associated with support for Harris among 2024 CES survey respondents.

The dataset is consistent with existing research; support for Democratic nominee Kamala Harris was higher among respondents with a favorable view of incumbent Democratic President Joe Biden and the economy [7, 8, 13]. As shown in Figure 4, 96.5% of respondents who believed that the economy had "gotten much better" supported Harris, while 86% of respondents who believed that the economy had "gotten much worse" supported Trump. Respondents who believed that the economy "stayed about the same" also favored Trump over Harris (57% vs 37.5%) [13]. Almost all respondents who "somewhat approve" or "strongly approve" of the Biden administration voted Harris, while over 90% of respondents who "strongly disapprove" of the Biden administration voted Trump [7, 13].

Race remains a significant predictor of vote choice, but its effects might vary across different regions, the urban-rural divide, and highest level of educational attainment [6]. Figure 5 shows the interaction of race and the rural-urban divide among respondents. White respondents living in cities overwhelmingly broke for Harris and white respondents living in rural areas overwhelmingly broke for Trump. On the contrary, Black respondents living in all types of areas favored Harris [13]. White male and female respondents had similar levels of support for Harris, but Black female respondents favored Harris more heavily than Black male respondents; this suggests that although support for Trump increased among racial minorities in 2024, white voters are still more likely to support Trump than voters of color [8, 11, 13].

4 Model

My modeling strategy aims to identify the logistic regression model that best fits the 2024 CES survey dataset, determine which of the predictors suggested by the works of Kuriwaki et al. and Algara et al. improve model fit, and train and optimize a logistic machine learning (ML) model to accurately classify Harris and Trump voters and generalize well. Logistic regression will be used to classify respondents by vote choice because it is a simple, interpretable model suitable for predicting this binary outcome of interest [8].

4.1 Frequentist model selection process

Likelihood ratio testing will be used to determine which of the models outlined below best explains presidential vote choice in the 2024 CES survey dataset and whether any of the race-related interaction predictors suggested by Kuriwaki et al or economy and Biden-approval related predictors suggested by Algara et al. can be ommitted [6] [7]. The null hypothesis will be that the reduced models fit the data as well as the full model [18]. Once the model of best fit has been determined, it will be trained and applied to the 2024 CES survey dataset and its performance will be reported. Bayesian model diagnostics are available in Appendix C

Model 0 (full model)

$$\begin{aligned} p(\text{vote\_trump} = 1) = \sigma\Big(&\beta_0 + \beta_1 \cdot \text{age\_bracket} + \beta_2 \cdot \text{gender} + \beta_3 \cdot \text{educ} + \beta_4 \cdot \text{state} \\ &+ \beta_5 \cdot \text{region} + \beta_6 \cdot \text{urbancity} + \beta_7 \cdot \text{biden\_approval} + \beta_8 \cdot \text{econ\_past\_year} \\ &+ \beta_9 \cdot \text{price\_change\_past\_year} + \beta_{10} \cdot \text{family\_income\_past\_year} + \beta_{11} \cdot (\text{race} \times \text{region}) \\ &+ \beta_{12} \cdot (\text{race} \times \text{urbancity}) + \beta_{13} \cdot (\text{race} \times \text{educ}) + \beta_{14} \cdot (\text{race} \times \text{gender})\Big) \end{aligned}$$

Model 1

$$\begin{aligned} p(\text{vote\_trump} = 1) = \sigma\Big(&\beta_0 + \beta_1 \cdot \text{age\_bracket} + \beta_2 \cdot \text{gender} + \beta_3 \cdot \text{educ} + \beta_4 \cdot \text{state} \\ &+ \beta_5 \cdot \text{region} + \beta_6 \cdot \text{urbancity} + \beta_7 \cdot \text{econ\_past\_year} + \beta_8 \cdot \text{price\_change\_past\_year} \\ &+ \beta_9 \cdot \text{family\_income\_past\_year} + \beta_{10} \cdot (\text{race} \times \text{region}) + \beta_{11} \cdot (\text{race} \times \text{urbancity}) \\ &+ \beta_{12} \cdot (\text{race} \times \text{educ}) + \beta_{13} \cdot (\text{race} \times \text{gender})\Big) \end{aligned}$$

Model 2

$$\begin{aligned} p(\text{vote\_trump} = 1) = \sigma\Big(&\beta_0 + \beta_1 \cdot \text{age\_bracket} + \beta_2 \cdot \text{gender} + \beta_3 \cdot \text{educ} + \beta_4 \cdot \text{state} \\ &+ \beta_5 \cdot \text{region} + \beta_6 \cdot \text{urbancity} + \beta_7 \cdot \text{biden\_approval} + \beta_8 \cdot (\text{race} \times \text{region}) \\ &+ \beta_9 \cdot (\text{race} \times \text{urbancity}) + \beta_{10} \cdot (\text{race} \times \text{educ}) + \beta_{11} \cdot (\text{race} \times \text{gender})\Big) \end{aligned}$$

Model 3

$$\begin{aligned} p(\text{vote\_trump} = 1) = \sigma\Big(&\beta_0 + \beta_1 \cdot \text{age\_bracket} + \beta_2 \cdot \text{gender} + \beta_3 \cdot \text{educ} + \beta_4 \cdot \text{state} \\ &+ \beta_5 \cdot \text{region} + \beta_6 \cdot \text{urbancity} + \beta_7 \cdot \text{biden\_approval} + \beta_8 \cdot \text{econ\_past\_year} \\ &+ \beta_9 \cdot \text{price\_change\_past\_year} + \beta_{10} \cdot \text{family\_income\_past\_year} + \beta_{11} \cdot (\text{race} \times \text{educ}) \\ &+ \beta_{12} \cdot (\text{race} \times \text{gender})\Big) \end{aligned}$$

Model 4

$$
\begin{aligned}
p(\text{vote\_trump} = 1) = \sigma\Big(&\beta_0 + \beta_1 \cdot \text{age\_bracket} + \beta_2 \cdot \text{gender} + \beta_3 \cdot \text{educ} + \beta_4 \cdot \text{state} \\
&+ \beta_5 \cdot \text{region} + \beta_6 \cdot \text{urbancity} + \beta_7 \cdot \text{biden\_approval} + \beta_8 \cdot \text{econ\_past\_year} \\
&+ \beta_9 \cdot \text{price\_change\_past\_year} + \beta_{10} \cdot \text{family\_income\_past\_year} + \beta_{11} \cdot (\text{race} \times \text{region}) \\
&+ \beta_{12} \cdot (\text{race} \times \text{urbancity})\Big)
\end{aligned}
$$

4.2 Selected model

The Python `statsmodels` package was used to perform likelihood ratio tests to determine whether the full model (Model 0) fits the 2024 CES survey data set and explains vote choice better than the reduced models (Models 1-4) [17] [18]. As shown in Table 1, The comparisons of Model 0 and Model 1 (LR Stat = 13,889.12) and by Model 0 vs Model 2 (LR Stat = 848.27) produced large likelihood ratio statistics relative to the comparisons of Model 0 and Model 3 (LR Stat = 48.15) and Model 0 vs Model 4 (LR Stat = 76.84). However all four likelihood ratio tests yielded p-values of less than 0.001 ($p < 0.001$); this indicates that the full model fits the data significantly better than each of the reduced models [19]. Incumbent presidential approval, economy-related varaibles, the interactions of race with the urban-rural divide and region, and the interactions of race with gender and highest level of educational attainment are all significant predictors of vote choice.

Reduced model	**Dummy predictors omitted**	**LR Stat**	**Df difference**	**p-value**
Model 1	biden_approval	13889.1241	4	≈ 0.0
Model 2	econ_past_year, price_change_past_year, family_income_past_year	848.2705	13	≈ 0.0
Model 3	race × urbancity, race × region	48.1516	21	0.000656
Model 4	race × educ, race × gender	76.8410	18	$3.0415 \cdot e^{-9}$

Table 1: Likelihood ratio test results comparing the fit of the full model (Model 0) to each of the reduced models (Model 1, Model 2, Model 3, Model 4). Model 0 vs Model 1 LRT has LR Stat=138,889.1241 and $p \approx 0.0 << 0.05$; Model 0 vs Model 2 LRT has LR Stat=848.27 and $p \approx 0.0 << 0.05$; Model 0 vs Model 3 LRT has LR Stat=48.1516 and $p = 0.000656 << 0.05$; Model 0 vs Model 4 LRT has LR Stat=76.841 and $p = 3.0415 \cdot e^{-9} << 0.05$. This indicates that the 4 binary indicator variables associated with biden_approval, 13 associated with econ_past_year, price_change_past_year, and family_income_past_year, 21 associated with race × urbancity and race × region, and 18 associated with race × educ and race × gender significantly improve model fit.

5 Results

5.1 2024 CES dataset

Model 0,

$$
\begin{aligned}
p(\texttt{vote\_trump} = 1) = \sigma\Big(&\beta_0 + \beta_1 \cdot \texttt{age\_bracket} + \beta_2 \cdot \texttt{gender} + \beta_3 \cdot \texttt{educ} + \beta_4 \cdot \texttt{state} \\
&+ \beta_5 \cdot \texttt{region} + \beta_6 \cdot \texttt{urbancity} + \beta_7 \cdot \texttt{biden\_approval} + \beta_8 \cdot \texttt{econ\_past\_year} \\
&+ \beta_9 \cdot \texttt{price\_change\_past\_year} + \beta_{10} \cdot \texttt{family\_income\_past\_year} + \beta_{11} \cdot (\texttt{race} \times \texttt{region}) \\
&+ \beta_{12} \cdot (\texttt{race} \times \texttt{urbancity}) + \beta_{13} \cdot (\texttt{race} \times \texttt{educ}) + \beta_{14} \cdot (\texttt{race} \times \texttt{gender})\Big)
\end{aligned}
$$

was trained on a subset of the 2024 CES survey dataset to classify Harris voters (negative class) and Trump voters (positive class). The dataset was split into a training dataset (random 75% subset) and a test dataset (random 25%

subset) using the `train_test_split` function from the `scikit-learn` package [20]. The logistic model architecture was built and compiled using the `Sequential` model from the `keras` package; it includes a linear layer, sigmoid activation layer, and an output layer [21]. Weights were be optimized via mini-batch gradient descent using `keras` and `TensorFlow` to minimize loss and maximize (validation) accuracy. L2-regularization was be used to guard against overfitting and discourage large weights [22]. Details about hyperparameter tuning, including selecting the number of epochs, batch size, learning rate, and L2-regularization penalty, can be found in Appendix B. Model 0 was applied to the validation dataset and predicted probabilities of support for Trump were converted into the vote_trump binary outcome variable using a threshold of 0.5 [8]. If, based on Model 0 estimates, the predicted probability of a respondent supporting Trump was at least 0.5, they were assigned to the positive class (vote_trump=1); otherwise, they were assigned to the negative class (vote_trump=0).

As displayed in Table 2, Model 0 achieved 92.8% validation accuracy on the 2024 CES survey dataset; it was able to correctly classify 92.8% of unseen data points (respondents) as Trump or Harris voters based on their features (age bracket, gender, state, region, whether they live in an urban or rural area, approval of the Biden administration, perception of the economy and price changes in the year leading up to the election, self-reported changes in income over the past year, and the interactions between their race and gender, race and urban city, and race and education).

Class	**Precision**	**Recall**	**F1-score**	**Support**
0 (Harris voters)	0.9428	0.9301	0.9364	6498
1 (Trump voters)	0.9092	0.9253	0.9171	4911
Test Accuracy	0.9280 (Total support: 11,409)			
Macro avg	0.9260	0.9277	0.9268	11,409
Weighted avg	0.9283	0.9280	0.9281	11,409

Table 2: Model performance on the validation dataset (25% unseen subset of the 2024 CES survey dataset) after training via mini-batch gradient descent with batch size 32, 100 epochs, learning rate $\alpha = 0.05$, and applying L2-regularization with penaty $\lambda = 0.001$.

For each class, the model had strong precision (% of predicted voters that were correctly classified), recall (% of actual voters that were correctly identified), and F1-score (weighted average of precision and recall) [20]. As displayed in Table 2, the model achieved 92.8% validation accuracy; 94.28% precision, 93.01% recall, and 93.64% F1-score on the negative class (Harris voters); 90.92% precision, 92.53% recall, and 91.71% F1-score on the positive class (Trump voters).

5.2 2020 CES dataset

Algara et al., and Camatarri applied their models to various U.S. election-related datasets [7, 8]. As an extension of the first iteration of this paper, the model was re-trained on and applied to the 2020 Congressional Election Study (CES) dataset to further evaluate its classification performance and ability to generalize [14]. Re-training the model on the 2020 CES dataset was necessary because Republican Donald Trump was the incumbent president at the time; consequently, incumbent presidential approval and positive perceptions of economic conditions would have been associated with support for him instead of the Democratic nominee [7, 12]. An identical 75% to 25% `train_test_split` in the `scikit-learn` package was used. Figure 8 shows that minibatch gradient descent training converged well before the 100th epoch, so the number of epochs was reduced to 50. All other hyper-parameters from optimizing on the 2024 CES dataset remained the same. The question about perceived changes in the price of everyday goods was not included in the 2020 CES survey, so the `price_change_past_year` variable was omitted; all other micro- and macro- level predictors from Model 0 were included.

The following model was trained on a subset of the 2020 CES dataset to classify Biden (negative class) and

Trump (positive class) voters:

$$
\begin{aligned}
p(\texttt{vote\_trump} = 1) = \sigma\Big(&\beta_0 + \beta_1 \cdot \texttt{age\_bracket} + \beta_2 \cdot \texttt{gender} + \beta_3 \cdot \texttt{educ} + \beta_4 \cdot \texttt{state} + \beta_5 \cdot \texttt{region} \\
&+ \beta_6 \cdot \texttt{urbancity} + \beta_7 \cdot \texttt{trump\_approval} + \beta_8 \cdot \texttt{econ\_past\_year} + \beta_9 \cdot \texttt{family\_income\_past\_year} \\
&+ \beta_{10} \cdot (\texttt{race} \times \texttt{region}) + \beta_{11} \cdot (\texttt{race} \times \texttt{urbancity}) + \beta_{12} \cdot (\texttt{race} \times \texttt{educ}) + \beta_{13} \cdot (\texttt{race} \times \texttt{gender})\Big)
\end{aligned}
$$

As shown in Table 3, the model achieved 96.52% accuracy, high precision (% of correctly-labeled data points), and high recall (% of actual positives identified) for the negative (Biden) and positive (Trump) classes on an unseen test subset of the 2020 CES data.

Class	Precision	Recall	F1-score	Support
0 (Biden voters)	0.9821	0.9593	0.9706	6536
1 (Trump voters)	0.9415	0.9740	0.9575	4392
Test Accuracy	0.9652 (Total support: 10,928)			
Macro avg	0.9618	0.9667	0.9640	10,928
Weighted avg	0.9658	0.9652	0.9653	10,928

Table 3: Model performance on an unseen (25% test) subset 2020 CES dataset after training via mini-batch gradient descent with batch size 32, 50 epochs, learning rate $\alpha = 0.05$, and applying L2-regularization with penaty $\lambda = 0.001$.

5.3 2016 CCES dataset

A comparable model integrating micro- and macro-level predictors to identify Trump and Clinton voters was applied to the 2016 CCES dataset [16] to assess generalizability and increase confidence in the overall modeling approach. Although neither the `urbancity` variable nor the information required to construct it were available in the 2016 CCES dataset, all other micro- and macro-level predictors were present. Questions regarding economic conditions vary in each iteration of the CES [14]; the economy-related predictors included in the 2016 model are the respondent's perception of the economy in the year leading up to the election and self-reported family income changes during Obama's second four-year term.

Following the approach outlined in Section 5.1, the model below was trained on a subset of the 2016 CCES dataset to classify Clinton (negative class) and Trump (positive class) voters:

$$
\begin{aligned}
p(\text{vote\_trump} = 1) = \sigma\Big(&\beta_0 + \beta_1 \cdot \texttt{age\_bracket} + \beta_2 \cdot \text{gender} + \beta_3 \cdot \texttt{educ} + \beta_4 \cdot \texttt{state} + \beta_5 \cdot \texttt{region} \\
&+ \beta_6 \cdot \texttt{obama\_approval} + \beta_7 \cdot \texttt{econ\_past\_year} + \beta_8 \cdot \texttt{income\_past\_4year} + \beta_9 \cdot (\texttt{race} \times \texttt{region}) \\
&+ \beta_{10} \cdot (\texttt{race} \times \texttt{educ}) + \beta_{11} \cdot (\texttt{race} \times \texttt{gender})\Big)
\end{aligned}
$$

As shown in Table 4, the model achieved 93.35% accuracy on an unseen test subset of the data.

Class	Precision	Recall	F1-score	Support
0 (Clinton voters)	0.9362	0.9413	0.9387	5438
1 (Trump voters)	0.9303	0.9242	0.9273	4607
Test Accuracy	0.9335 (Total support: 10,045)			
Macro avg	0.9332	0.9328	0.9330	10,045
Weighted avg	0.9335	0.9335	0.9335	10,045

Table 4: Model performance on an unseen (25% test) subset of the 2016 CCES dataset after training via mini-batch gradient descent with batch size 32, 50 epochs, learning rate $\alpha = 0.05$, and applying L2-regularization with penaty $\lambda = 0.001$.

6 Discussion and Conclusion

6.1 Consistency with existing literature

The results presented in Section 4.2 provided strong evidence against the the null hypothesis, which stated that reduced models that omitted Biden approval, economy-related, interactions between race and demographic variables, or interactions between race and geographic variables would fit the data as well as the full model. The removal of the biden_approval term (Model 0 vs Model 1) resulted in a degrees of freedom difference of only 4, but the largest likelihood ratio statistic (LR Stat = 13,889.27). The Model 0 vs Model 1 (LR Stat = 13,889.12) and Model 0 vs Model 2 (LR Stat = 848.27) likelihood ratio tests yielded large likelihood ratio statistics compared to Model 0 vs Model 3 (LR Stat = 48.16) and Model 0 vs Model 4 (LR Stat - 76.84). This means that incumbent presidential (Biden) approval had the strongest effect on vote choice, followed by the combined effect of all three economy-related variables. The combined effect of the interactions between race and region and race and the urban-rural divide and the combined effect of the interactions between race and gender and race and highest level of educational attainment still significantly improved model fit, though the significance of these predictors is low relative to incumbent presidential approval and economic variables. This is consistent with the findings of Algara et al., who found that incumbent presidential approval is a key predictor of incumbent party vote share and predicted that Harris would narrowly lose due to Biden's low approval and Abramowitz, who found that incumbent presidential approval and a strong economy make individuals more likely to vote for the incumbent party's nominee [7] [12].

6.2 Machine learning model performance

The logistic vote choice models achieved 92.8%, 96.52%. and 93.35% test accuracy on the 2024, 2020, and 2016 CES datasets. This demonstrates that the modeling strategy outlined in Section 4 is generalizable and that a logistic vote choice model incorporating macro-level factors such as perceptions of economic conditions and incumbent presidential approval and micro-level sociodemographic variables including the interactions of race with gender, education, region, and the urban-rural divide can accurately classify Democratic and Republican presidential voters on unseen data. The results presented in Section 5 are consistent with Camatarri's suggestion that individual-level logistic models that incorporate micro- and macro-level predictors have the potential to accurately forecast aggregate election outcomes [8]. The results serve as a necessary first step towards an aggregated forecast. An aggregate popular vote or electoral college estimate based on logistic regression voter-level models is impossible if the logistic model cannot accurately classify voters.

6.3 Future work

Future work might involve exploring downsampling, Bayesian sampling methods, or incorporating the provided survey weights to address the mild class imbalances in the data and applying the models to voter registration or census data to forecast a future election. The majority class (Clinton/Biden/Harris) voters are over-represented and the minority class (Trump voters) is under-represented in each dataset. The CES sample was weighted to match American Communities Survey (ACS) distributions on age, gender, race, Hispanic origin, and highest level of educational attainment; weights were then post-stratified by age, gender, race, highest level of educational attainment, and past presidential votes [13]. Each dataset includes post-stratification weights, with higher weights assigned to respondents from strata that are underrepresented in the survey or less likely to respond to a survey [13, 15, 16]. Sampling methodologies were not used to address the class imbalance in this project because it is a mild imbalance (minority class represents more than 40% of each dataset) and it does not appear to significantly impact model performance or induce model bias towards the positive class [23]. As shown in Section 5, there is no significant precision, recall, or F1-score gap between the positive and negative class on any of the included datasets. However, it might be beneficial to re-train the models on balanced versions of the datasets (where Clinton/Biden/Harris voters are randomly downsampled) because training a model on an imbalance dataset risks bias towards the majority class and biased performance on unseen data [23]. The end goal to combine Frequentist, Machine Learning, and Bayesian methods to forecast the 2028 U.S. Presidential Election. This project is a first-step towards that; logistic models that can accurately classify unseen 2016, 2020, and 2024 CES survey respondents as Clinton/Harris/Biden or Trump voters were built, trained, and optimized. The project explores logistic individual-level vote choice modeling, reviews existing research, and builds a foundation for incorporating multiple methodological approaches in a future election forecast.

References

[1] CNN. ”Election 2024: Presidential Results,” CNN Politics, Election Day in America, 2024. Available: https://www.cnn.com/election/2024/results/president?election-data-id=2024-PG&election-painting-mode=projection-with-lead&filter-key-races=false&filter-flipped=false&filter-remaining=false

[2] P. Mongrain and M. Stegmaier, “Introduction to Forecasting the 2024 US Elections,” PS: Political Science & Politics, vol. 58, no. 2, pp. 184–191, 2025. doi:10.1017/S1049096524001008

[3] T. Hunnicutt, N. Bose, and S. Kelly. ”Kamala Harris made a historic dash for the White House. Here’s why she fell short.” Reuters, 2024. Available: https://www.reuters.com/investigations/kamala-harris-made-historic-dash-white-house-heres-why-she-fell-short-2024-11-06/

[4] M. Brenan. ”Economy Most Important Issue to 2024 Presidential Vote,” Gallup Politics, 2024. Available: https://news.gallup.com/poll/651719/economy-important-issue-2024-presidential-vote.aspx

[5] Pew Research Center, “In tied presidential race, Harris and Trump have contrasting strengths, weaknesses,” Pew Research Center: Politics, 2024. [Online]. Available: https://www.pewresearch.org/politics/2024/09/09/in-tied-presidential-race-harris-and-trump -have-contrasting-strengths-weaknesses/.

[6] S. Kuriwaki, J. Grumbach, A. Sadin, and E. Schickler, “The Geography of Racially Polarized Voting: Calibrating Surveys at the District Level,” Institution for Social and Policy Studies, Yale University, ISPS22-28, 2023. [Online]. Available: https://isps.yale.edu/research/publications/23-28

[7] C. Algara, L. Gomez, E. Headington, H. Liu, and B. Nigri, “Forecasting Partisan Collective Accountability during the 2024 US Presidential and Congressional Elections,” PS: Political Science & Politics, vol. 58, no. 2, pp. 211–218, 2025. doi:10.1017/S1049096524000854

[8] S. Camatarri, “Predicting Popular-vote Shares in US Presidential Elections: A Model-based Strategy Relying on Anes Data,” PS: Political Science & Politics, vol. 58, no. 2, pp. 253–257, 2025. doi:10.1017/S1049096524000933

[9] The American Presidency Project, ”2024,” Data Archive, Elections, 2024. Available at: https://www.presidency.ucsb.edu/statistics/elections/2024

[10] Python Software Foundation, Python, version 3.10, Python Software Foundation, 2021. [Online]. Available: https://www.python.org

[11] H. Hartig, S. Keeter, A. Daniller, and T. Van Green, “Behind Trump’s 2024 Victory, a More Racially and Ethnically Diverse Voter Coalition,” Pew Research Center, Jun. 26, 2025. [Online]. Available: https://www.pewresearch.org/politics/2025/06/26/behind-trumps-2024-victory-a-more-racially-and-ethnically-diverse-voter-coalition

[12] A. I. Abramowitz, “An Improved Model for Predicting Presidential Election Outcomes,” *PS: Political Science and Politics*, vol. 21, no. 4, pp. 843–847, 1988. doi: 10.2307/420023.

[13] B. Schaffner, M. Shih, S. Ansolabehere, and J. Pope, “Cooperative Election Study Common Content, 2024.” Harvard Dataverse, 2025. doi: doi:10.7910/DVN/X11EP6.

[14] J. M. Tisch College of Civic Life, ”Cooperative Election Study,” Tufts University, 2025. [Online]. Available: https://tischcollege.tufts.edu/research-faculty/research-centers/cooperative-election-study

[15] S. Ansolabehere, B. Schaffner, and S. Luks, ”CCES Common Content, 2020,” Harvard Dataverse, version 4.0, 2020. [Online]. Available: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FE9N6PH

[16] S. Ansolabehere, B. Schaffner, and S. Luks, ”CCES Common Content, 2016,” Harvard Dataverse, 2017. [Online]. Available: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FGDF6Z0

[17] S. Seabold and J. Perktold, “Statsmodels: Statistical Modeling and Econometrics in Python.” [Online]. Available: https://www.statsmodels.org/stable/index.html

[18] Wikipedia, “Likelihood-ratio test.” [Online]. Available: https://en.wikipedia.org/wiki/Likelihood-ratio_test

[19] UCLA Statistical Consulting Group, ”How are the likelihood ratio, Wald, and Lagrange multiplier (score) tests different and/or similar?” *UCLA: Statistical Consulting Group*, [Online]. Available: https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faqhow-are-the-likelihood-ratio-wald-and-lagrange-multiplier-score-tests-different-andor-similar/

[20] F. Pedregosa et al., “3.4 Metrics and Scoring: quantifying the quality of predictions,” scikit-learn: Machine Learning in Python, scikit-learn developers, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html

[21] Keras Team, “keras,” PyPI, version 3.11.1, Jul. 31, 2025. [Online]. Available: https://pypi.org/project/keras/

[22] Google Developers. ”Overfitting: L2 regularization,” Machine Learning: ML Concepts, 2024. [Online] Available: https://developers.google.com/machine-learning/crash-course/overfitting/regularization

[23] Google Developers. ”Datasets, generalization, and overfitting: Imbalanced datasets,” Machine Learning: ML Concepts, 2024. [Online] Available: https://developers.google.com/machine-learning/crash-course/overfitting/imbalanced-datasets

[24] R. Fonnesbeck, “PyMC: Bayesian Modeling in Python,” PyMC Documentation, [Online]. Available: https://www.pymc.io/welcome.html

[25] R. Kumar, “ArviZ: Exploratory Analysis of Bayesian Models,” ArviZ Documentation, [Online]. Available: https://python.arviz.org/en/stable/.

[26] R. Gopinath “Bayesian MMM: Diagnostics, Analysis, and Comparison.” Medium, 2024. [Online]. Available: https://medium.com/@mail2rajivgopinath/bayesian-mmm-diagnostics-analysis-and-comparison-6a34e727c543

A Additional data details

A.1 Data preparation

The preliminary release of the 2024 CES survey dataset was downloaded from Harvard Dataverse via url using the `pandas` `read_stata` function. It was published on April 2, 2025 [13]. The raw dataset contains 60,000 rows (responses from American adults) and 684 columns [13]. The raw 2024 CES dataset was subset to exclude respondents who did not participate in the post-election wave due to a lack of information about who they voted for in the 2024 presidential race and respondents with a NULL response to one of the included questions to allow for data to be converted into dummies using the `pandas` package and subsequently used to fit and train models using `statsmodels`, `TensorFlow`, and `keras`.

The following columns from the raw dataset set were used in my analysis:

- CC24_364a: who the respondent voted for in the 2024 U.S. Presidential Election
- gender4: the respondent's gender identity (male, female, non-binary, or other)
- birthyr: the year that the respondent was born
- race: the respondent's self-reported race
- hispanic: whether the respondent self-identifies as Hispanic
- inputstate: the respondent's state of residence
- region: the region of the U.S. that the respondent lives in
- urbancity: type of area that the respondent lives in
- CC24_301: the respondent's perception of how the national economy changed in the year leading up to the election
- CC24_302: self-reported change in respondent's family income in the year leading up to the election
- CC24_303: respondent's perception of how the prices of everyday goods changed in the year leading up to the election

CC24_364a was used to construct the vote_trump binary indicator variable, birthyr was used to construct the age_bracket categorical variable, and the dataset was subsetted to only include respondents who voted for Harris or Trump and identify as either male or female prior to fitting logistic models. This was done to simplify this analysis into a classification task for logistic regression and to prevent sparse gender and race $\times$ gender categories. Following the data cleaning methodology of Kuriwaki et al., I included the following categories for race to prevent sparse race and race-related interaction categories: Black, White, Hispanic, and All Other Races [6]. Sparse categories cause errors when using `statsmodels.api` to fit logistic models and perform likelihood ratio testing [17].

For the final iteration of this paper, analogous data cleaning steps were performed on the 2016 CCES [16] and 2020 CES [15] survey datasets.

A.2 Visualizations

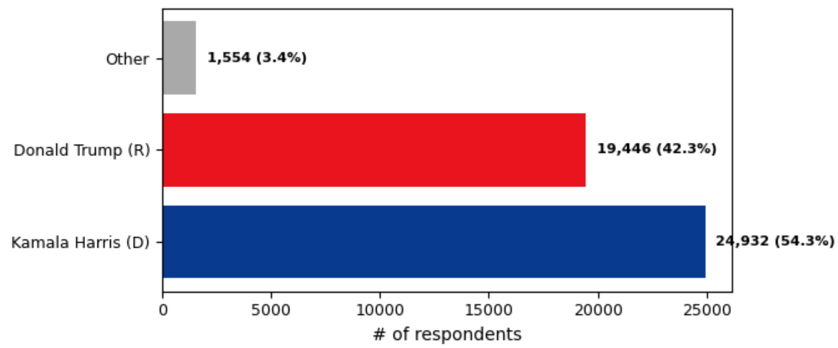

Figure 1: Among 2024 CES respondents, 54.3% voted for Harris, 42.3% voted for Trump, and 3.4% voted third-party or write-in in the 2024 Presidential Election.

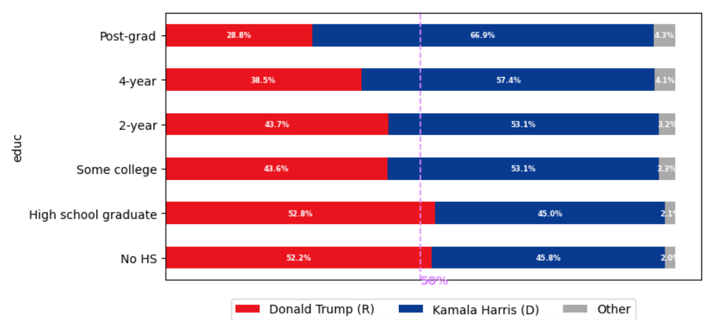

Figure 2: A majority of 2024 CES respondents with at least some college supported Harris, with highest Harris support among respondents with a post-graduate (advanced) degree. A majority of respondents without a college education supported Trump.

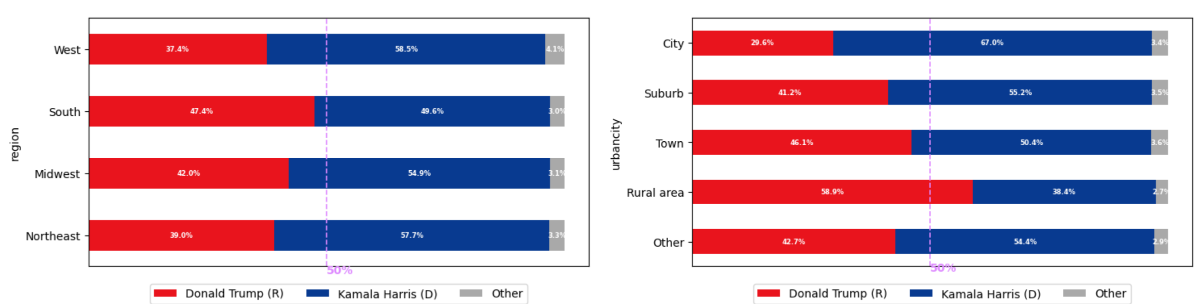

Figure 3: Support for Harris was higher among 2024 CES respondents who live in the West, the Northeast, and in urban areas (cities and suburbs). Support for Trump was higher among survey respondents who live in the South or in rural areas/small towns.

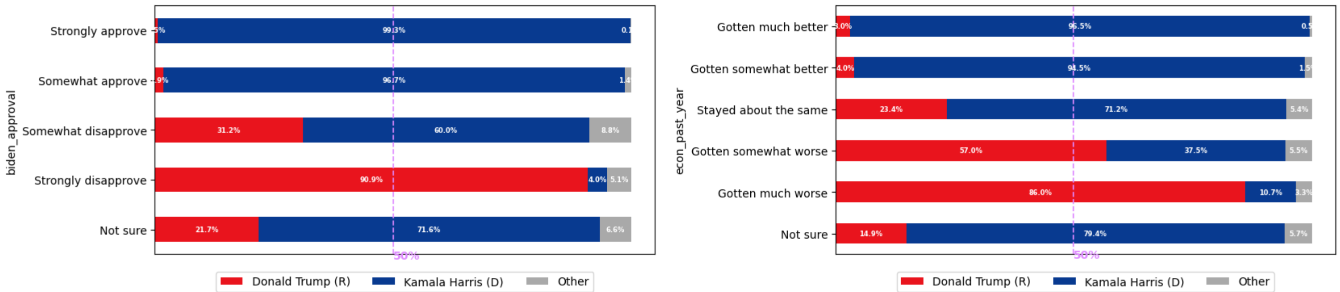

Figure 4: Positive perceptions of the state of the economy in the year leading up to the 2024 presidential election and approval of the Biden administration are associated with support for Harris, while negative perceptions and disapproval of the Biden administration are associated with support for Trump among 2024 CES respondents.

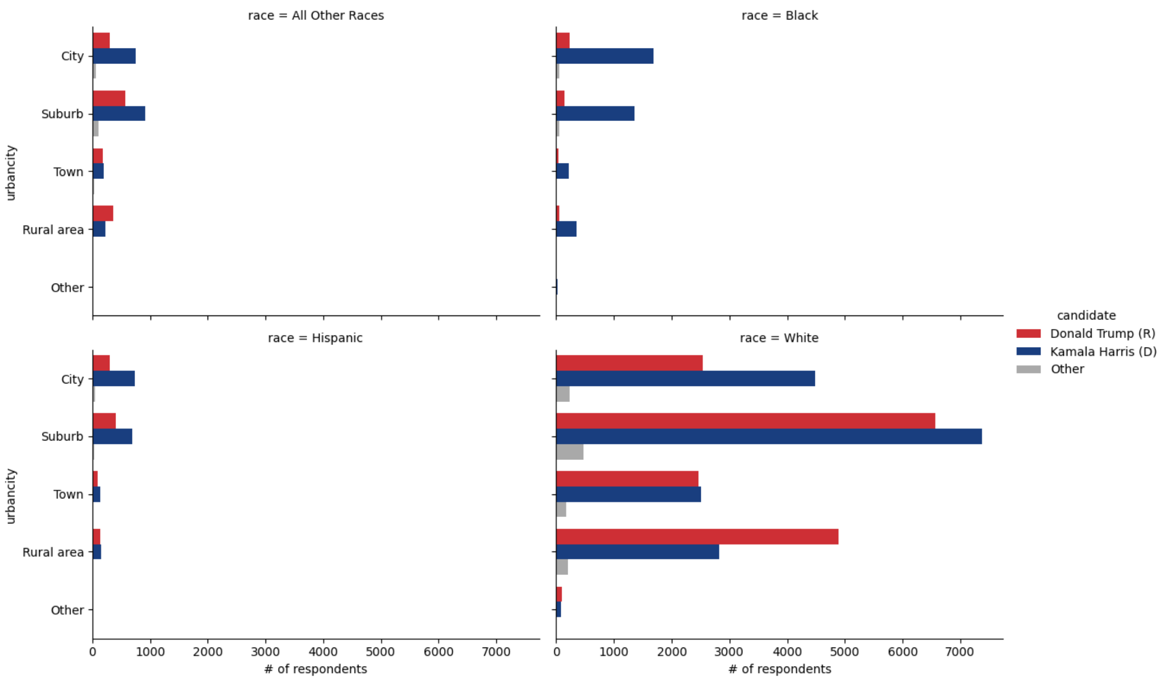

Figure 5: Presidential vote breakdown among respondents, by race and type of area among 2024 CES respondents.

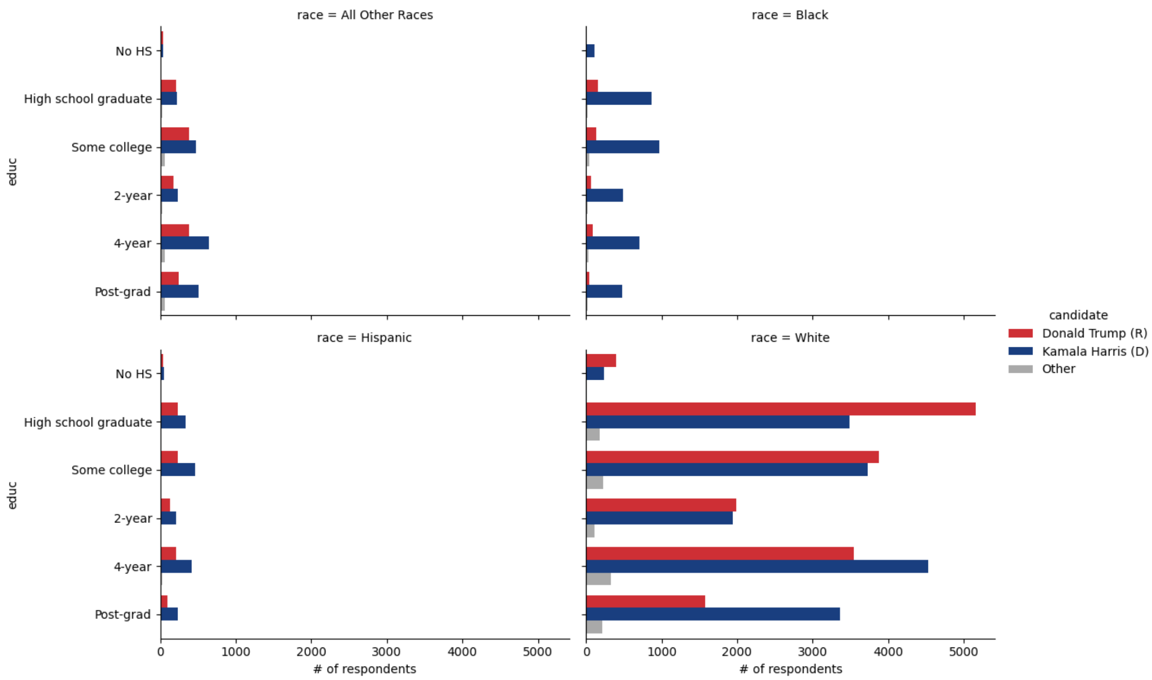

Figure 6: Presidential vote breakdown among respondents, by race and highest level of educational attainment among 2024 CES respondents.

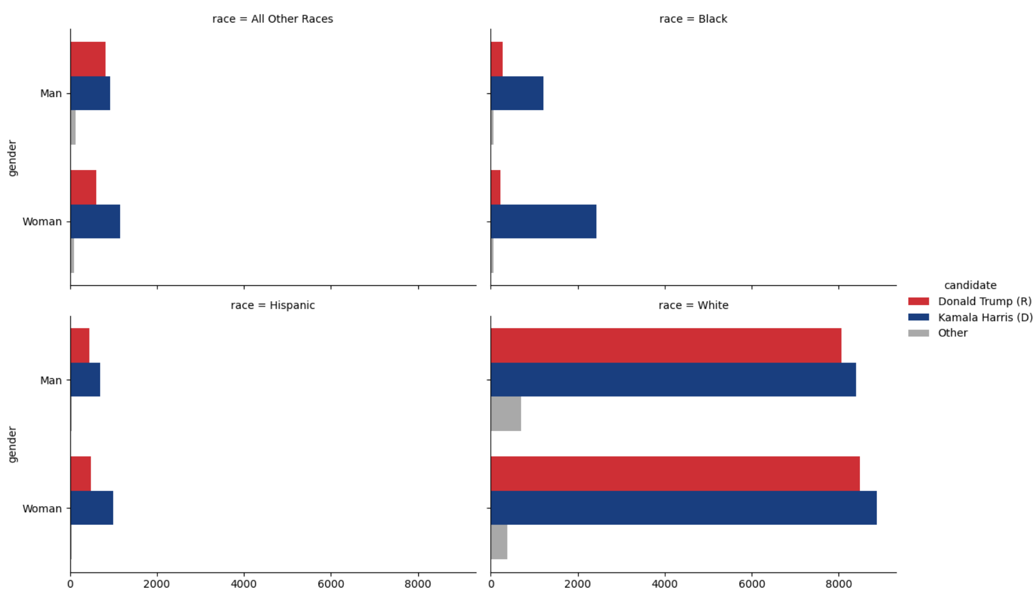

Figure 7: Presidential vote breakdown among 2024 CES respondents, by race and gender.

B Model training and optimization

B.1 Hyperparameter tuning

The TensorFlow and keras Python libraries were used to perform mini-batch gradient descent and tune hyperparameters for the logistic vote choice machine learning model outlined in Section 5.1. Hyperparameter tuning was done only on the 2024 CES dataset.

Batch Size	Learning Rate	Training loss	Validation accuracy
32	0.001	0.3117	88.27%
	0.01	0.2244	92.24%
	0.05	0.1971	92.68%
	0.1	0.1951	92.75%
64	0.0001	0.4561	80.6%
	0.01	0.2248	92.26%
	0.05	0.2046	92.52%
	0.1	0.1977	92.61%
128	0.0001	0.6384	75.21%
	0.001	0.2768	91.10%
	0.05	0.2123	92.42%
	0.1	0.2083	92.57%
256	0.001	1.0418	58.50%
	0.01	0.3008	89.22%
	0.05	0.2324	92.07%
	0.1	0.2126	92.55%

Table 5: Training loss and validation accuracy at different batch sizes and learning rates after 100 epochs of mini-batch gradient descent.

λ	Training cost	Validation accuracy
0.001	0.2024	92.71%
0.01	0.2168	92.77%
0.1	0.2859	92.76%
1	0.6277	58.48%

Table 6: Training loss and validation accuracy for different values of λ with batch size 32, learning rate $\alpha = 0.05$, and 100 epochs. The model without regularization achieved 92.68% accuracy with the same number of epochs, batch size, and learning rate. Regularization improved validation accuracy by approximately one-tenth of a percentage point.

B.2 Loss and accuracy versus number of epochs using batch size 32, learning rate $\alpha = 0.05$, and L2 penalty $\lambda = 0.001$.

Figure 8 shows loss and accuracy versus number of epochs using optimized hyperparameters from the previous section. It appears that the gradient descent algorithm converges before 50 epochs; this suggests that early stopping or reducing the number of epochs might be an appropriate next step.

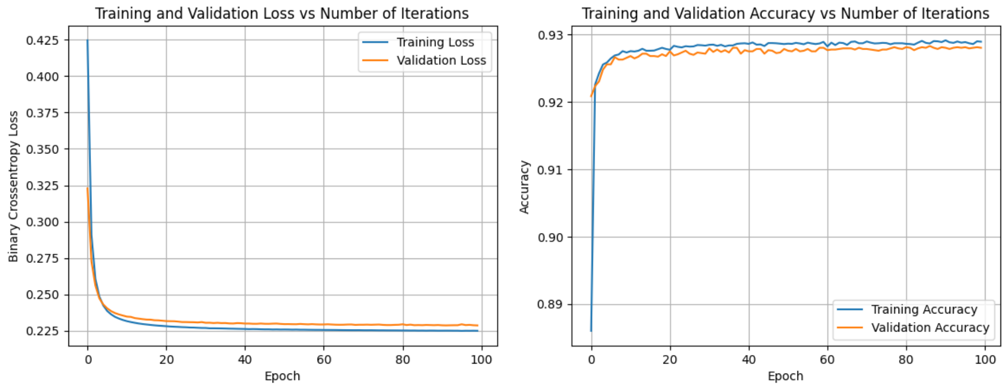

Figure 8: Training and Validation Loss and Training and Validation Accuracy on the 2024 CES dataset; 100 epochs, batch size = 32, learning rate α = 0.05, and L2 penalty λ = 0.001.

C Model diagnostics

Model checks and diagnostics for Model 0 (outlined in 5.1) were performed using PyMC [24] on the 2024 CES dataset. Arviz was used to report Bayesian model analysis, including trace plots, R-hat plots, and posterior predictive checks [25] [26]. The posterior predictive checks displayed in Figure 9 show that model-generated data (inference data) is closely aligned with the observed data (2024 CES survey responses) [25] [26]. Figure 10 displays the trace plots for the intercept and weights. The trace plots show how the values of each weight change over time, with random fluctuations around a constant mean, indicating that the model has converged [26]. Figure 11 shows R-hat values very close to 1 and smaller than 1.01 for all predictors [25]. This indicates that the chains mixed well and converged [24] [26]. Figure 12 shows the posterior distribution of each model parameter [25].

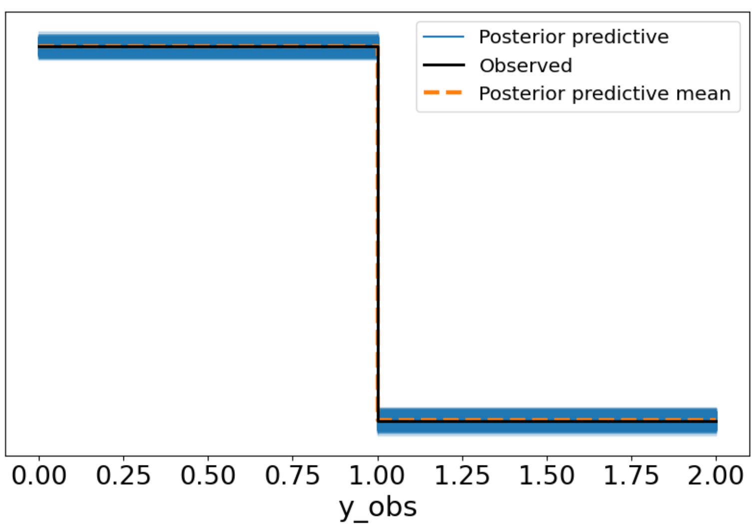

Figure 9: Posterior predictive checks

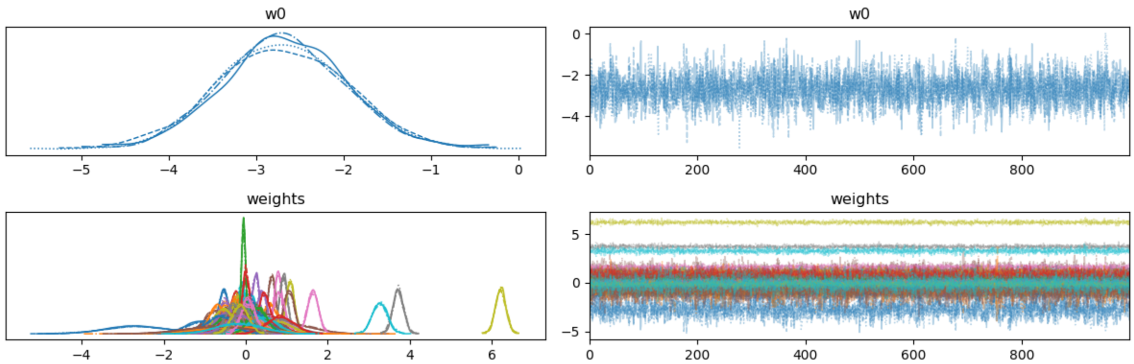

Figure 10: Trace plots

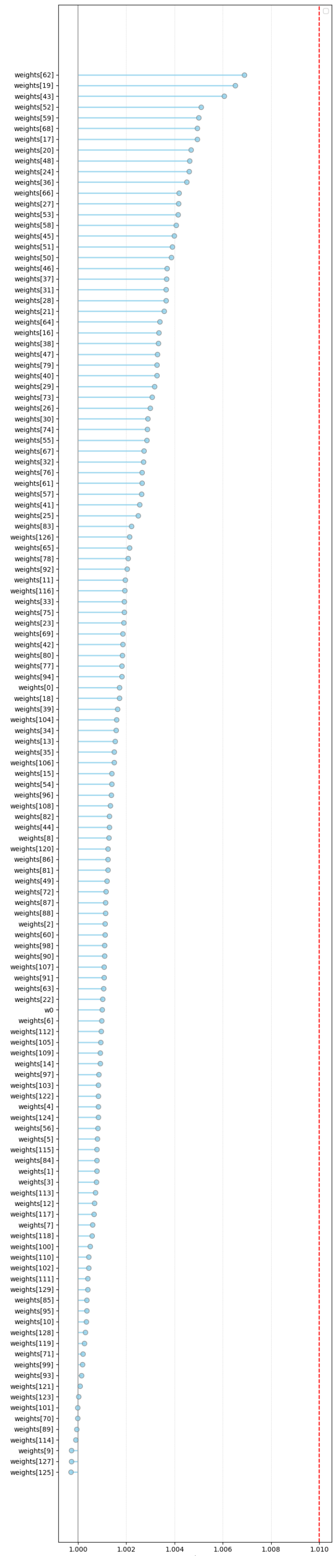

Figure 11: R-hat plots

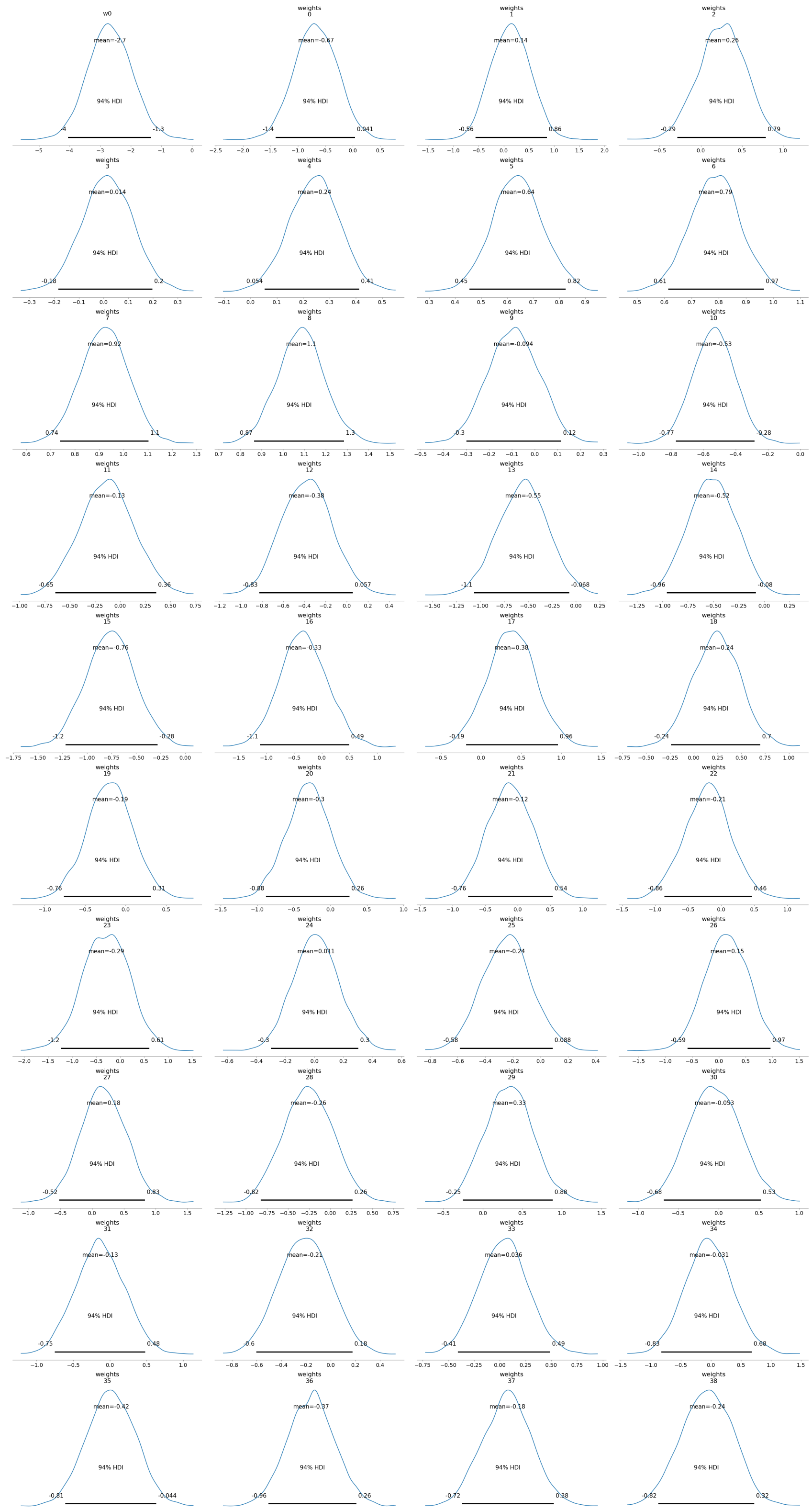

Figure 12: Posterior plots

D Extension to 2016 CCES and 2020 CES Datasets

D.1 Loss and accuracy versus number of epochs using batch size 32, learning rate $\alpha = 0.05$, and L2 penalty $\lambda = 0.001$

The same hyperparameters shown in Appendix B, with the exception of the number of epochs, were used when the model was re-trained on and applied to the 2016 CCES and the 2020 CES datasets. The 2024 CES dataset training curves shown in Figure 8 suggest that the gradient converged before 100 epochs, so the number of epochs was reduced to 50. The batch size, learning rate, and L2-penalty from Appendix B were re-used.

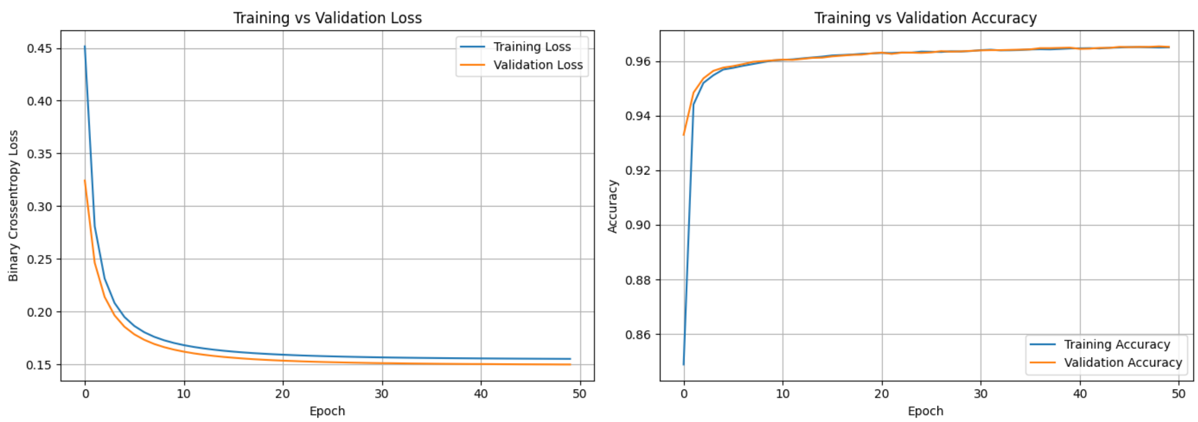

Figure 13: Training and Validation Loss and Training and Validation Accuracy vs Number of minibatch gradient descent iterations (epochs) on the 2020 CES dataset; 50 epochs, batch size = 32, learning rate $\alpha = 0.05$, and L2 penalty $\lambda = 0.001$.

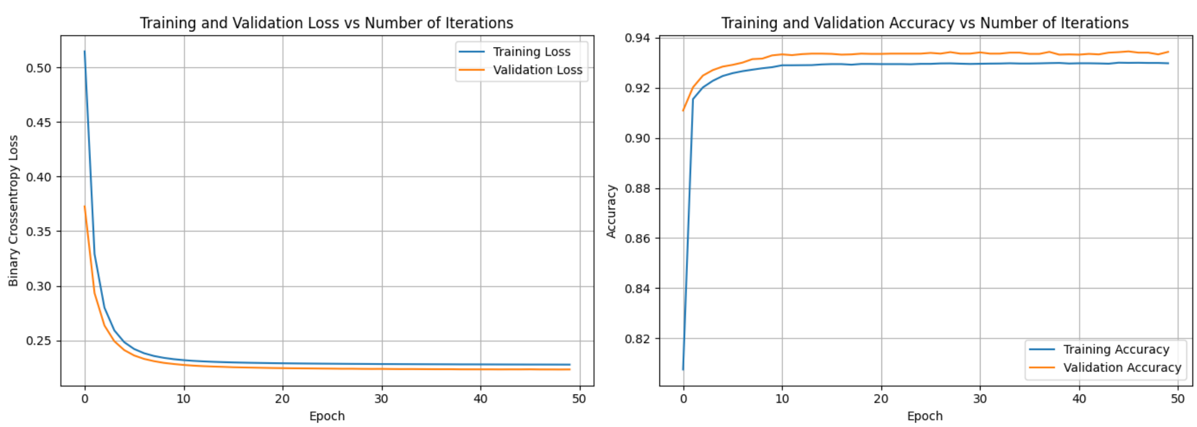

Figure 14: Training and Validation Loss and Training and Validation Accuracy on the 2016 CCES dataset; 50 epochs, batch size = 32, learning rate $\alpha = 0.05$, and L2 penalty $\lambda = 0.001$.