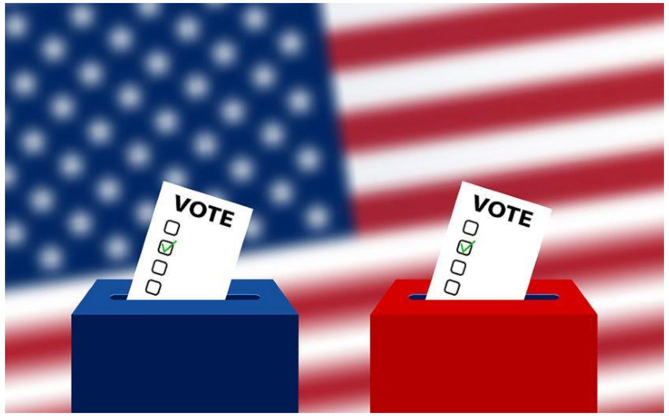
Logistic Regression Analysis of Individual-Level Vote Choice in Recent U.S. Presidential Elections



Talia Fabregas

University of Toronto Department of Statistical Sciences

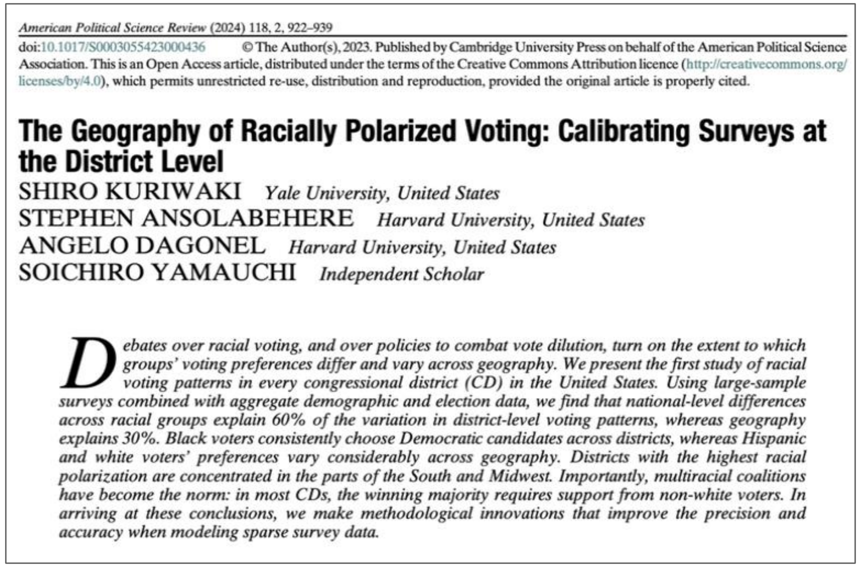
STA496 Summer Research Presentation

Supervisor: Professor Rohan Alexander

August 2025

Overview

- ***How do micro-*** *(race, gender, age, education, geography, race-related interactions, and other socioeconomic variables)* ***and macro-*** *(incumbent presidential approval and economic conditions)* ***level factors influence individual vote choice in U.S. Presidential elections?***
- This project builds on the works of **Kuriwaki et al. (2023),** who found that **race is a significant predictor** of vote choice, but its **interactions with education and geography must also be considered**; **Algara et al. (2024),** who found that **incumbent presidential approval and economic condition**s are significant predictors of incumbent party vote share;, and Camatarri (2024) who proposes the use of **micro- and macro-level predictors** in standard and Bayesian logistic models to **predict aggregate election outcomes**.



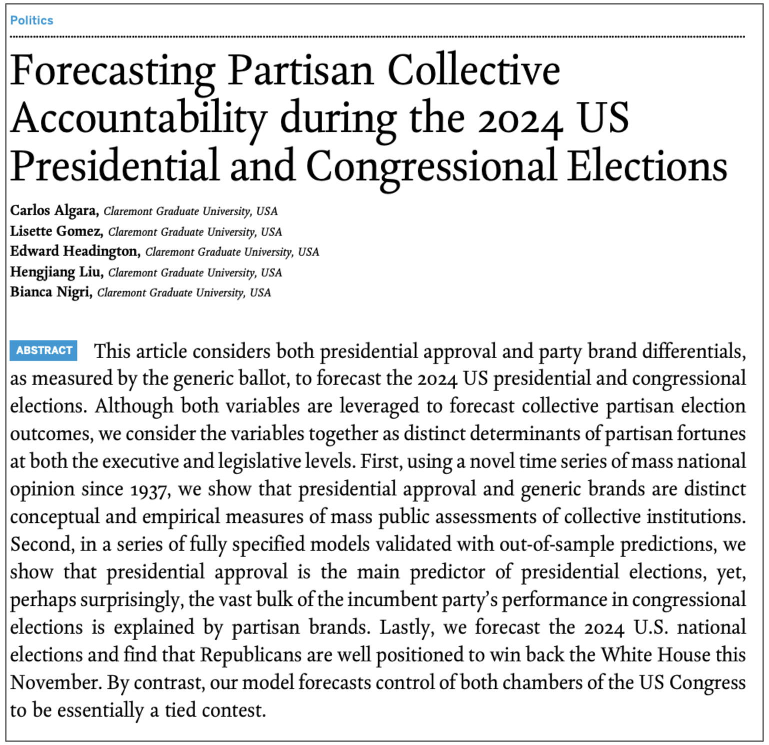
American Political Science Review (2024) 118, 2, 922–939

doi:10.1017/S0003055423000436

The Geography of Racially Polarized Voting: Calibrating Surveys at the District Level

SHIRO KURIWAKI *Yale University, United States*
STEPHEN ANSOLABEHERE *Harvard University, United States*
ANGELO DAGONEL *Harvard University, United States*
SOICHIRO YAMAUCHI *Independent Scholar*

Debates over racial voting, and over policies to combat vote dilution, turn on the extent to which groups' voting preferences differ and vary across geography. We present the first study of racial voting patterns in every congressional district (CD) in the United States. Using large-sample surveys combined with aggregate demographic and election data, we find that national-level differences across racial groups explain 60% of the variation in district-level voting patterns, whereas geography explains 30%. Black voters consistently choose Democratic candidates across districts, whereas Hispanic and white voters' preferences vary considerably across geography. Districts with the highest racial polarization are concentrated in the parts of the South and Midwest. Importantly, multiracial coalitions have become the norm: in most CDs, the winning majority requires support from non-white voters. In arriving at these conclusions, we make methodological innovations that improve the precision and accuracy when modeling sparse survey data.

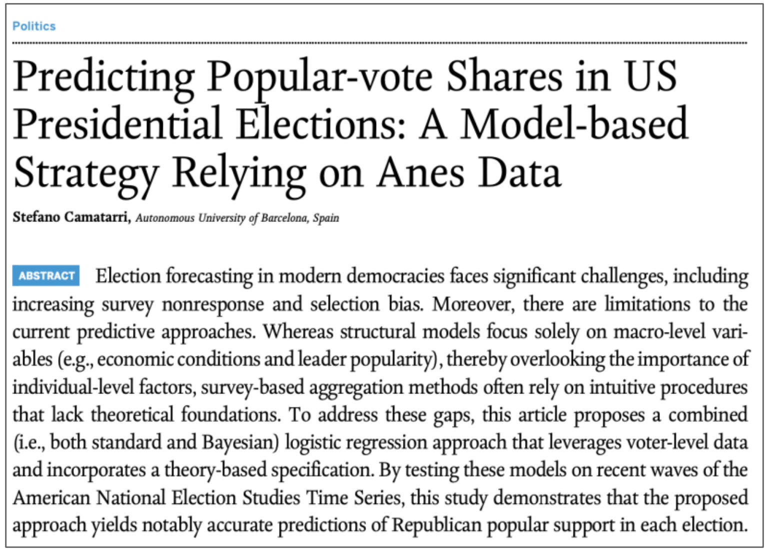


Politics

Forecasting Partisan Collective Accountability during the 2024 US Presidential and Congressional Elections

Carlos Algara, *Claremont Graduate University, USA*
Lisette Gomez, *Claremont Graduate University, USA*
Edward Headington, *Claremont Graduate University, USA*
Hengjiang Liu, *Claremont Graduate University, USA*
Bianca Nigri, *Claremont Graduate University, USA*

ABSTRACT This article considers both presidential approval and party brand differentials, as measured by the generic ballot, to forecast the 2024 US presidential and congressional elections. Although both variables are leveraged to forecast collective partisan election outcomes, we consider the variables together as distinct determinants of partisan fortunes at both the executive and legislative levels. First, using a novel time series of mass national opinion since 1937, we show that presidential approval and generic brands are distinct conceptual and empirical measures of mass public assessments of collective institutions. Second, in a series of fully specified models validated with out-of-sample predictions, we show that presidential approval is the main predictor of presidential elections, yet, perhaps surprisingly, the vast bulk of the incumbent party's performance in congressional elections is explained by partisan brands. Lastly, we forecast the 2024 U.S. national elections and find that Republicans are well positioned to win back the White House this November. By contrast, our model forecasts control of both chambers of the US Congress to be essentially a tied contest.



Politics

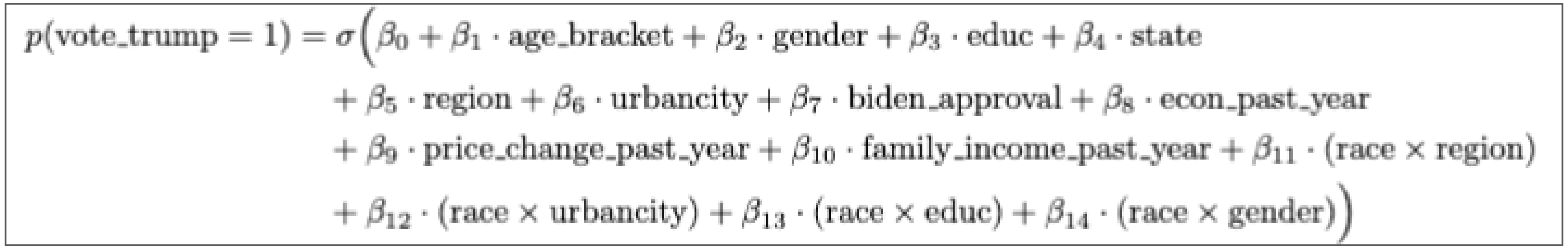
Predicting Popular-vote Shares in US Presidential Elections: A Model-based Strategy Relying on Anes Data

Stefano Camatarri, *Autonomous University of Barcelona, Spain*

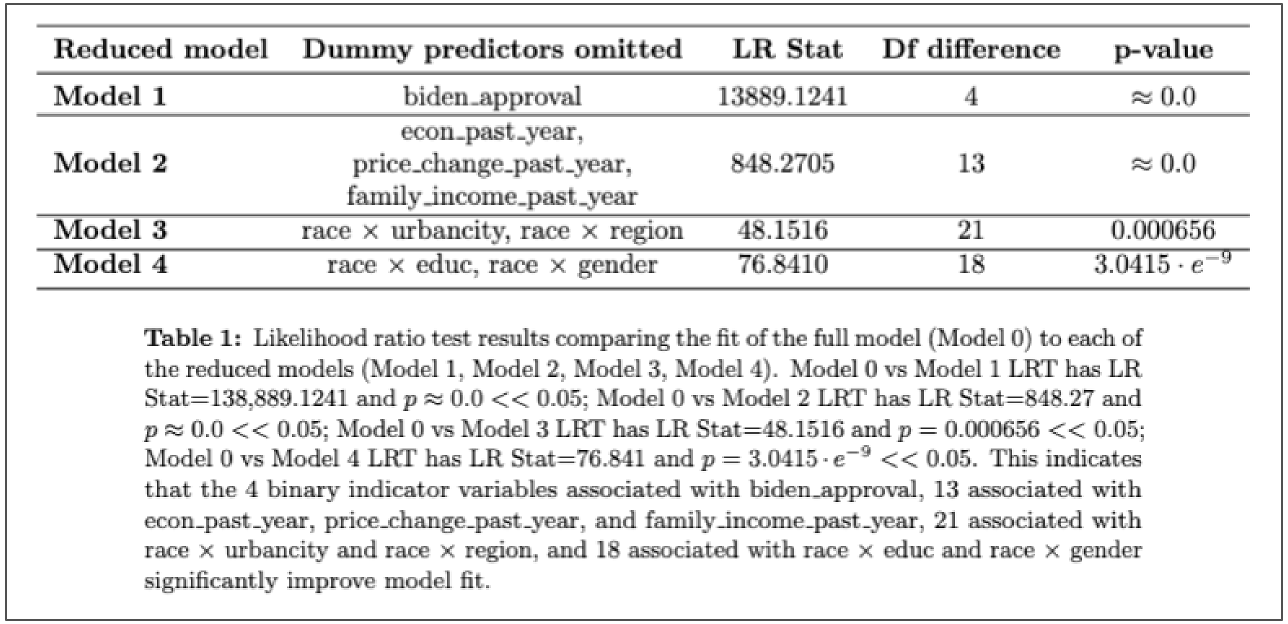
ABSTRACT Election forecasting in modern democracies faces significant challenges, including increasing survey nonresponse and selection bias. Moreover, there are limitations to the current predictive approaches. Whereas structural models focus solely on macro-level variables (e.g., economic conditions and leader popularity), thereby overlooking the importance of individual-level factors, survey-based aggregation methods often rely on intuitive procedures that lack theoretical foundations. To address these gaps, this article proposes a combined (i.e., both standard and Bayesian) logistic regression approach that leverages voter-level data and incorporates a theory-based specification. By testing these models on recent waves of the American National Election Studies Time Series, this study demonstrates that the proposed approach yields notably accurate predictions of Republican popular support in each election.

Model Selection

This project aims to build and train a logistic regression model that uses micro- and macro-level predictors to classify supporters of the Democratic and Republican presidential nominees using the Python `scipy, scikit-learn`, `TensorFlow`, and `Keras` frameworks. Drawing inspiration from Camatarri (2024), I approach vote choice modeling as a binary classification problem and fit the following standard logistic regression model on the 2024 Cooperative Election Study (CES) dataset and used likelihood ratio testing for feature selection.



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



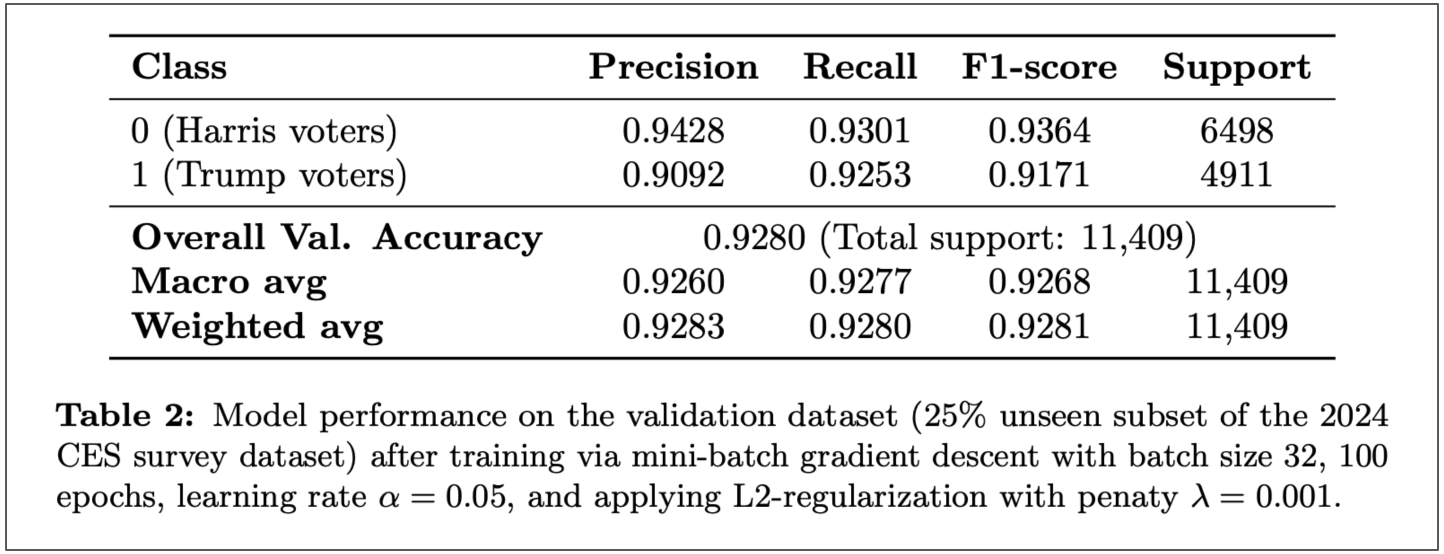
Reduced model	Dummy predictors omitted	LR Stat	Df difference	p-value
Model 1	biden_approval	13889.1241	4	≈ 0.0
Model 2	econ_past_year, price_change_past_year, family_income_past_year	848.2705	13	≈ 0.0
Model 3	race × urbancity, race × region	48.1516	21	0.000656
Model 4	race × educ, race × gender	76.8410	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.

Results

2024 Cooperative Election Study (CES) dataset

- After performing frequentist feature selection, I trained and optimized the full logistic model on the 2024 CES dataset using `sci-kit learn`, `TensorFlow`, and `keras`.
- **92.8% test accuracy** on a random 25% unseen subset of the data (default 75%:25% scikit-learn train test split); high precision, recall, and F1-scores for the positive (Trump) and negative (Harris) class).



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
Overall Val. 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 $\lambda = 0.001$.

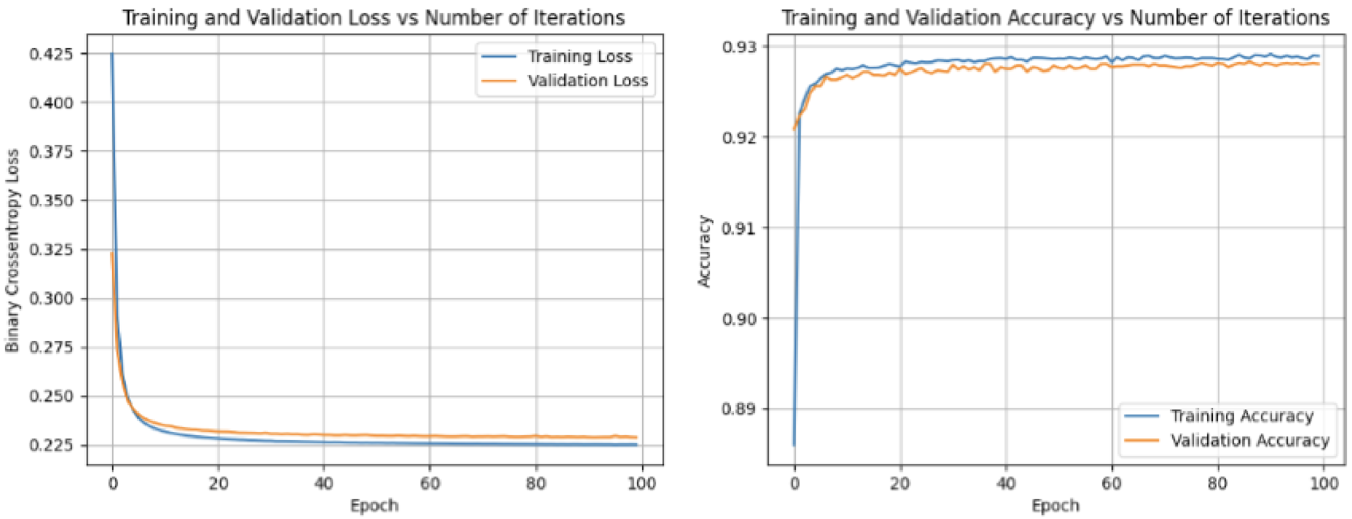
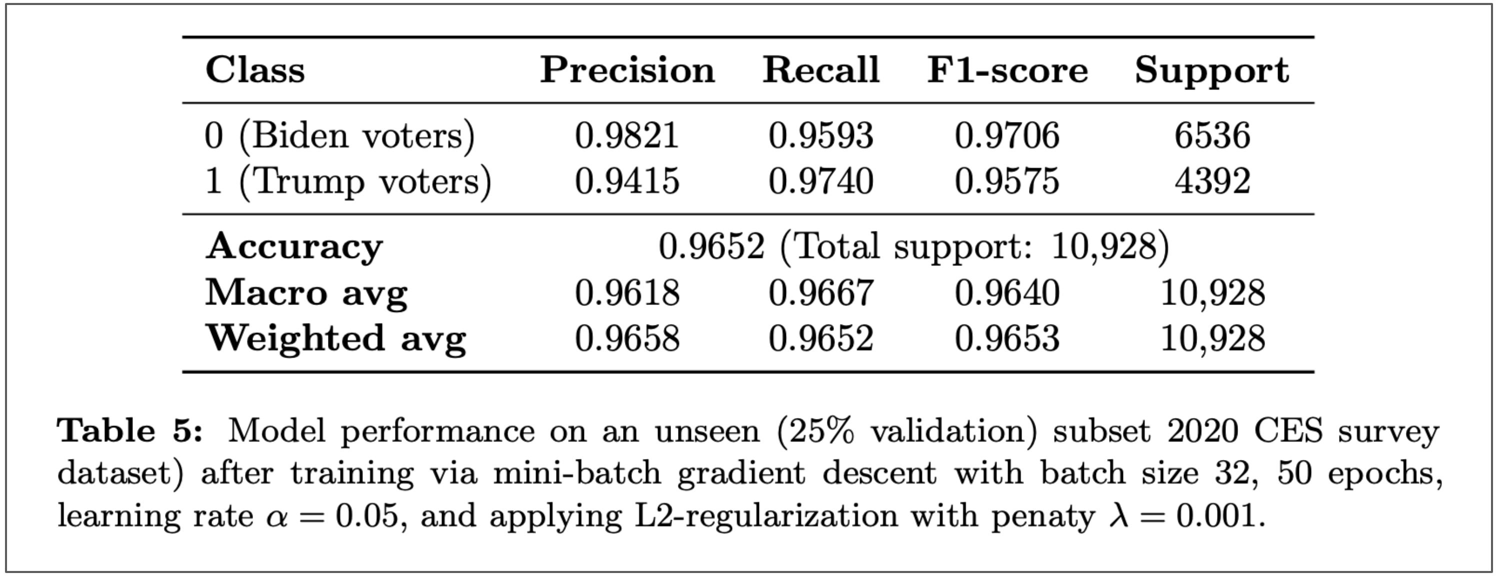


Figure 8: Training and Validation Loss and Training and Validation Accuracy on the 2024 CES dataset; 100 epochs, batch size = 32, learning rate $\alpha = 0.05$, and L2 penalty $\lambda = 0.001$.

2020 Cooperative Election Study (CES) dataset

- As an extension, the model was re-trained and tested on the 2020 Congressional Election Study (CES) dataset using the same predictors, except for `price_change_past_year` (question not included in 2020 survey).

- **Achieved higher (96.52%) test accuracy** on 25% unseen subset of 2020 CES data; default 75%:25% scikit-learn train-test split was used. Number of epochs was reduced to 50; no other hyperparameters were changed.



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
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 5: Model performance on an unseen (25% validation) subset 2020 CES survey 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 $\lambda = 0.001$.

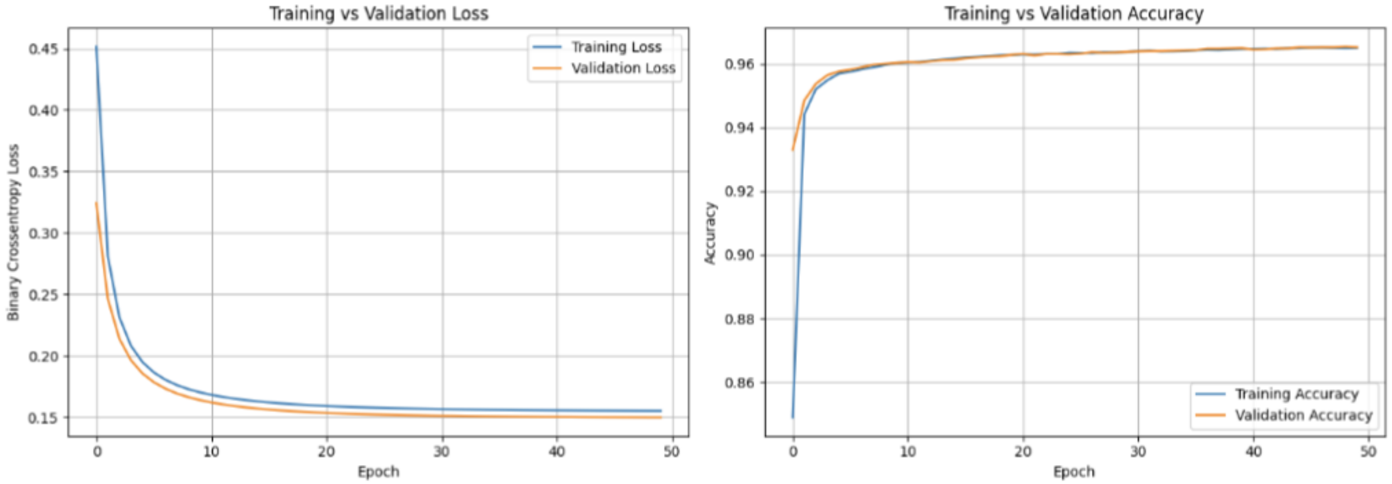


Figure 13: Training and Validation Loss and Training and Validation Accuracy on the 2020 CES dataset; 50 epochs, batch size = 32, learning rate $\alpha = 0.05$, and L2 penalty $\lambda = 0.001$.

Conclusion & Takeaways

- Removal of incumbent presidential approval predictors had the most detrimental effect on model fit (Kuriwaki et al. 2023).
- A logistic vote choice model that combines macro-level predictors (economic conditions, incumbent presidential approval) with micro-level predictors (age, gender, race, education, geography and other individual-level sociodemographic variables) achieved strong predictive performance on two survey datasets.
- Logistic models that can predict individual voting behavior have the potential to produce accurate aggregated election forecasts (Camatarri 2024).

My project repo can be found via this link:

References

Algara, C., Gomez, L., Headington, E., Liu, H., & Nigri, B. (2025). Forecasting Partisan Collective Accountability during the 2024 US Presidential and Congressional Elections. *PS: Political Science & Politics*, *58*(2), 211–218. doi:10.1017/S1049096524000854

KURIWAKI, S., ANSOLABEHERE, S., DAGONEL, A., & YAMAUCHI, S. (2024). The Geography of Racially Polarized Voting: Calibrating Surveys at the District Level. *American Political Science Review*, *118*(2), 922–939. doi:10.1017/S0003055423000436

Camatarri, S. (2025). Predicting Popular-vote Shares in US Presidential Elections: A Model-based Strategy Relying on Anes Data. *PS: Political Science & Politics*, *58*(2), 253–257. doi:10.1017/S1049096524000933

Note: Full project reference list is available at the end of my paper titled “Trump (again): Logistic Regression Analysis of Individual-Level Vote Choice in the 2024 Presidential Election”.