2020 National Election Day Exit Poll

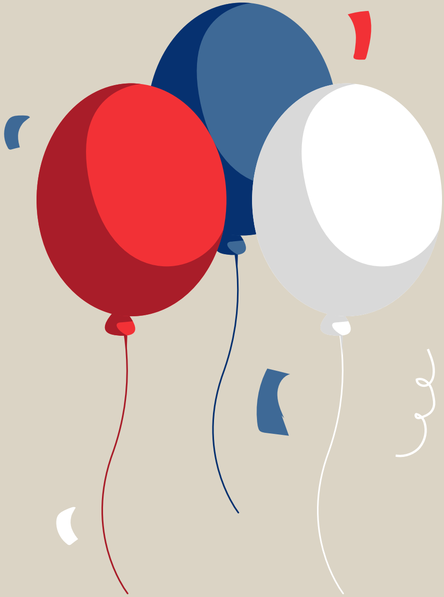
Statistics 112 - Group 7

Oliver Siu, Tiffany Hu, Jasmine Jungreis,
Christopher Le, & Dillon Maheshwari

Abstract

Our analysis aimed to explore patterns in voting behavior using demographic and social attitudes data. From EDA and random forest, we identified variables such as higher education levels, support for legal abortion, and belief in climate change, as positively associated with voting Democratic, while better financial situations and Christian religious affiliations were negatively associated. We discovered the effectiveness of a combined reduced model (abortion stance, climate change stance, financial situation, religion, education and race) through comparison of prediction accuracy and ROC AUC scores.

TABLE OF CONTENTS



1 Question of Interest

2 EDA

3 Model Selection

4 Model Assumptions

5 Results

6 Conclusion

1

Question of Interest

Research Question
Variable Selection
Schematic

Research Question

Can we predict the likelihood of a person voting Democratic or Republican based on their personal demographic information and social attitude?

Data Resource

Dataset: National Election Pool Poll: 2020 National Election Day Exit Poll

The dataset consists of survey data collected from 2020 Exit Poll Surveys for the U.S.

Participants' responses were recorded via telephone or in-person interviews outside voting centers.

The dataset aggregates 4 survey versions that were distributed throughout the U.S.

National Election Pool (ABC News, CBS, CNN, NBC). (2020). National Election Pool Poll: 2020 National Election Day Exit Poll (Version 4) [Dataset]. Roper Center for Public Opinion Research. doi:10.25940/ROPER-31119913

Variables

Original Dataset:
Observations: 15,351
Variables: 118

Cleaned Dataset:
Observations: 3,163
Variables: 16

Survey Version 4 was chosen because it contained most of the variables relevant to our research question.

Variables of interest were further narrowed down based on what we thought would be most influential.

All variables are categorical.

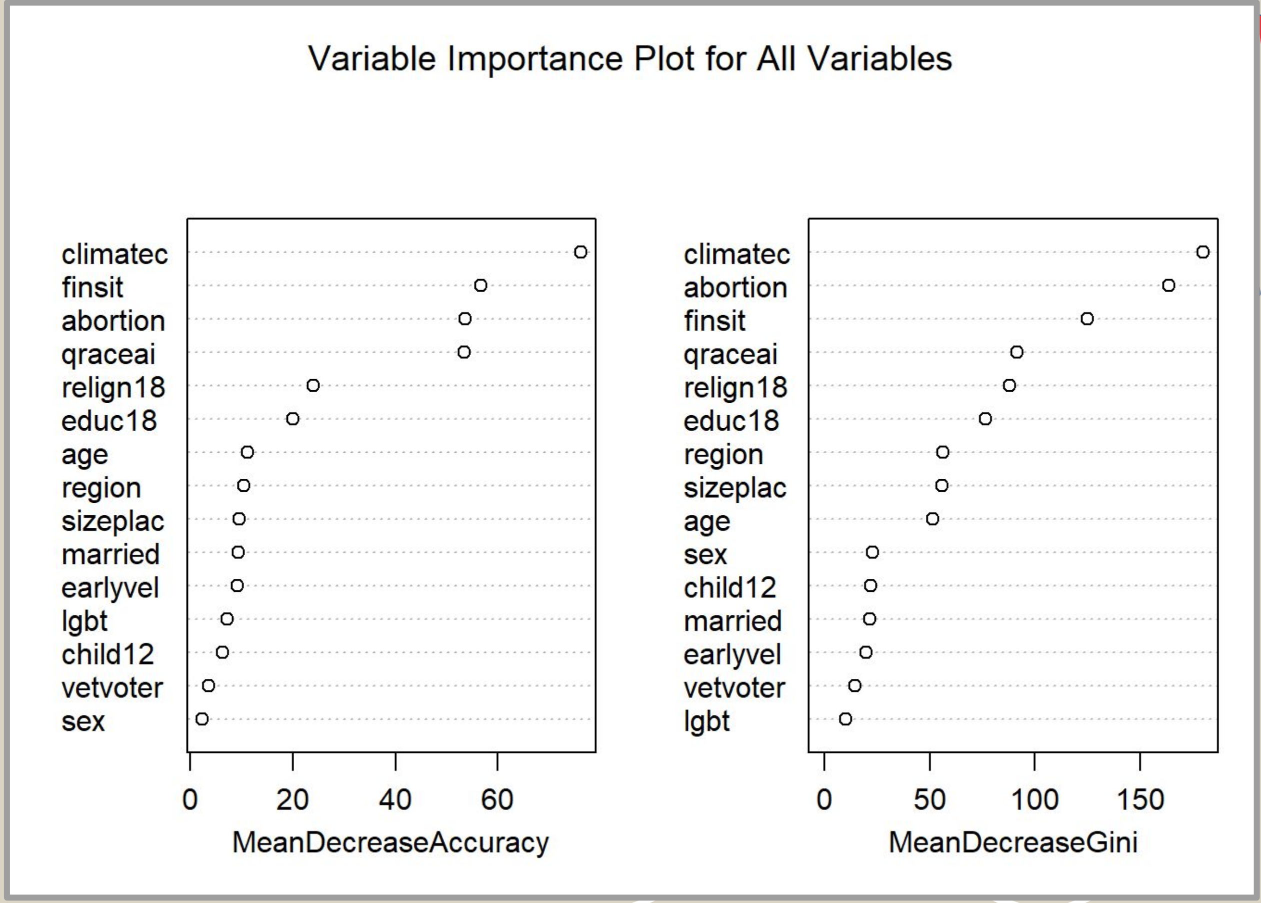
Variable	Description / Survey Question	Levels
pres	President interviewee voted for (Joe Biden or Donald Trump)	2
age	Interviewee’s age group	4
educ18	Interviewee’s education level	5
earlyvel	Voter type (in-person or other)	2
qraceai	Interviewee’s race	6
region	Region where interviewee resides (North, South, East, or West)	4
sex	Interviewee’s sex	2
sizeplac	Population of the area where the interviewee resides	5
abortion	Interviewee’s stance on abortion	5
climatec	Do you think climate change, also known as global warming, is a serious problem?	3
lgbt	Are you gay, lesbian, bisexual or transgender?	2
child12	Do you have any children under 18 living in your household?	2
finsit	Interviewee’s household financial situation	4
married	Are you currently married?	2
relign18	Interviewee’s religion	7
vetvoter	Have you ever served in the U.S. military?	2

Variable Importance

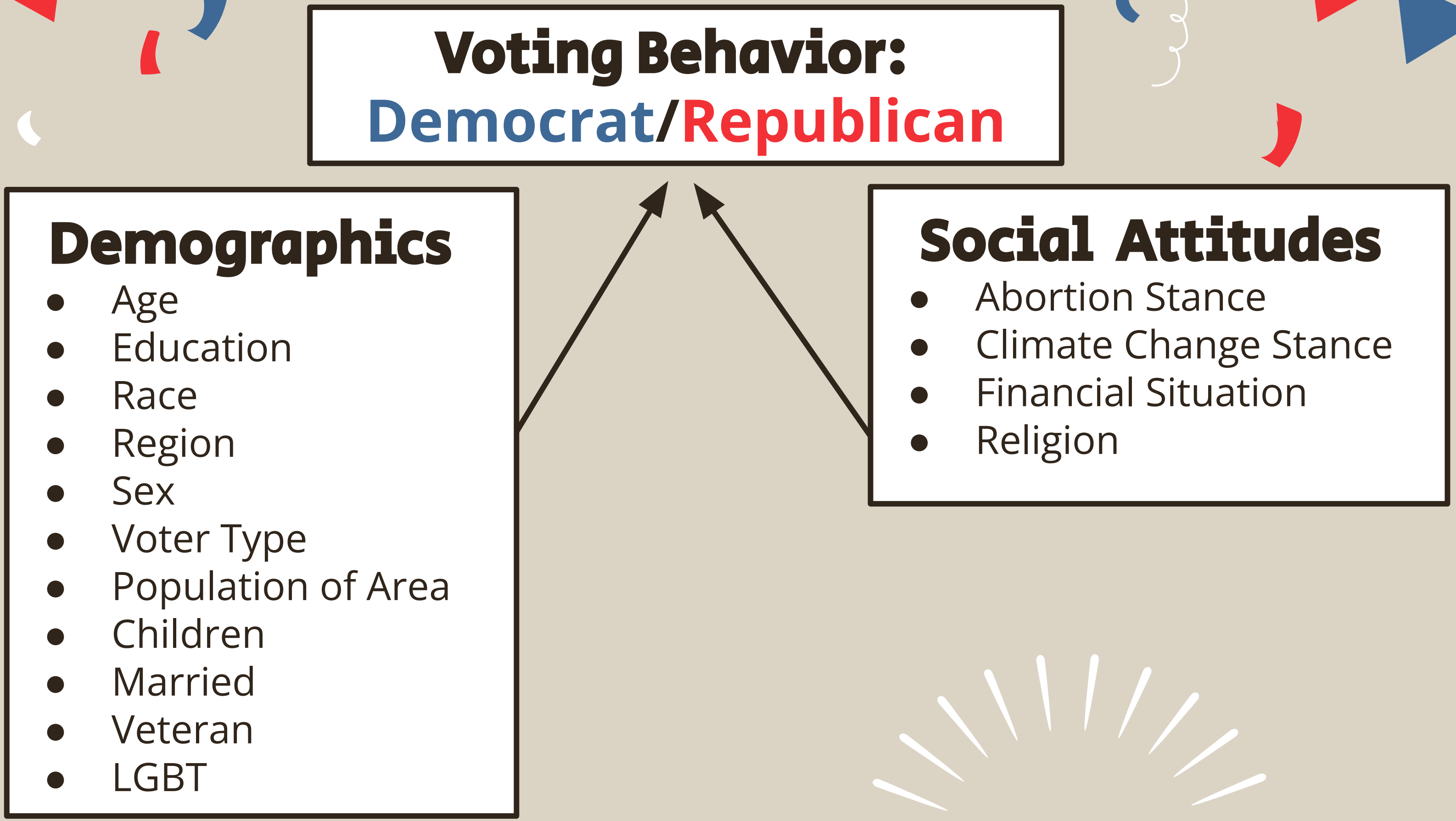
Random Forest was used to aid variable selection.

The top 6 most influential predictors are:

educ18: **Education Level**
qraceai: **Race**
abortion: **Abortion Stance**
climatec: **Climate Change Stance**
finsit: **Financial Situation**
relign18: **Religion**



Schematic



Voting Behavior:
Democrat/Republican

Demographics

- Age
- Education
- Race
- Region
- Sex
- Voter Type
- Population of Area
- Children
- Married
- Veteran
- LGBT

Social Attitudes

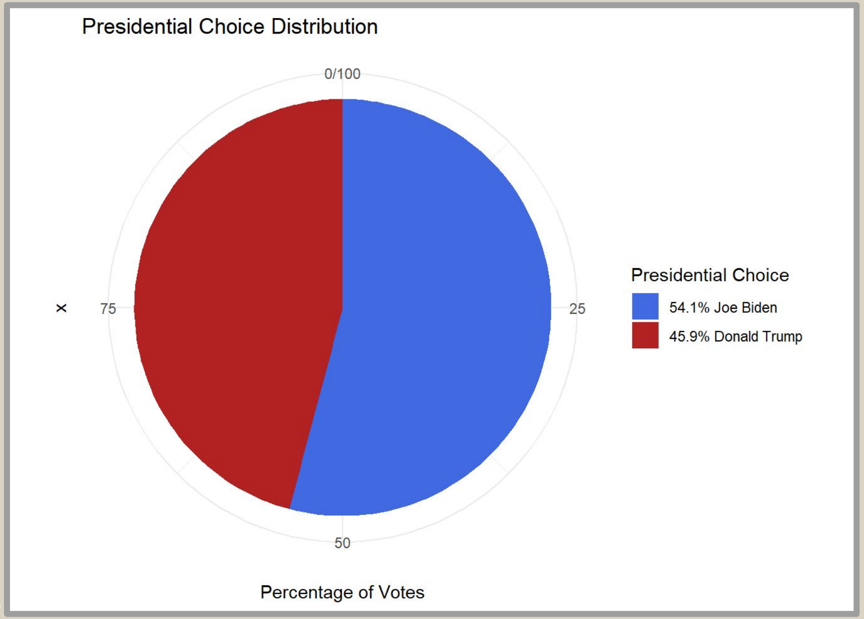
- Abortion Stance
- Climate Change Stance
- Financial Situation
- Religion

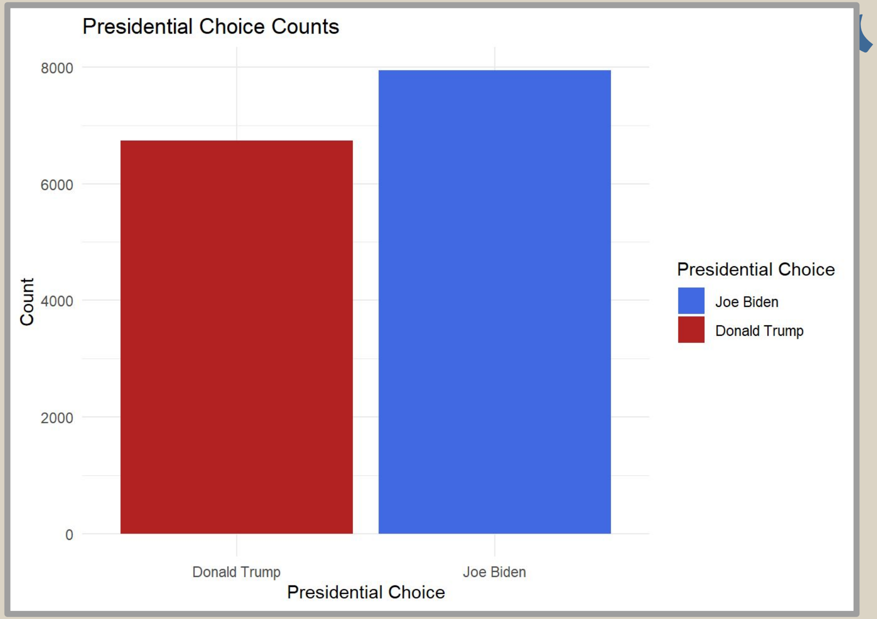
2

EDA

Exploratory Data Analysis

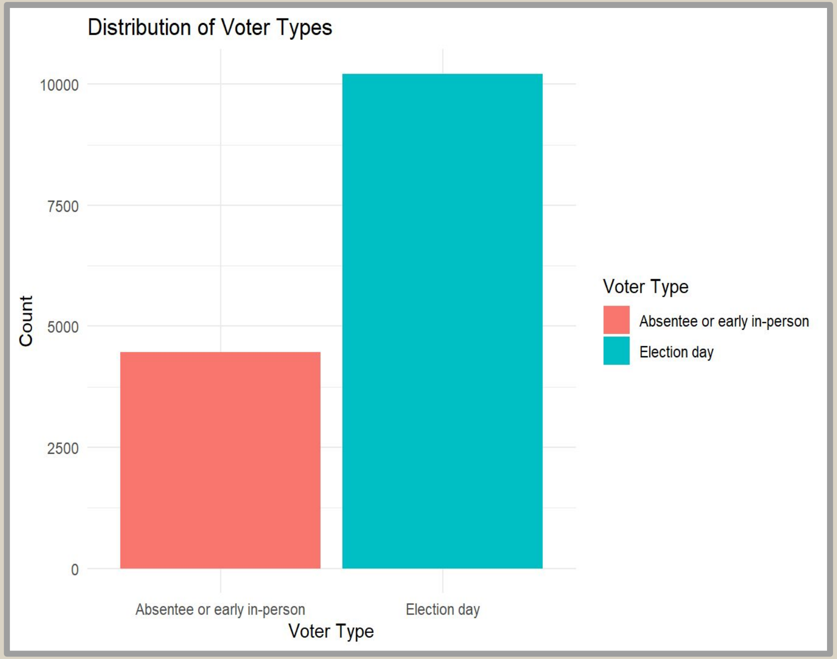
Presidential Choice Distribution

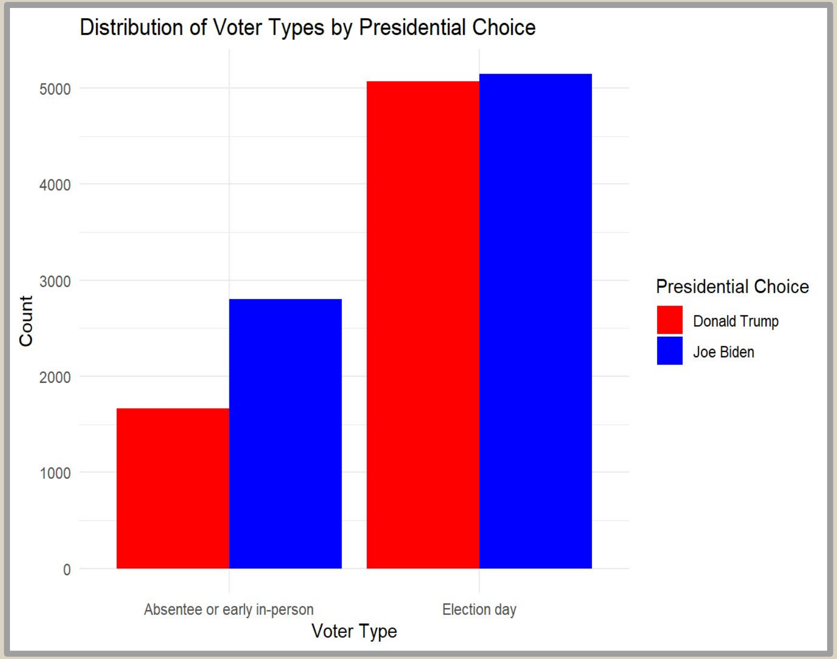




- In the 2020 United States presidential election, Joe Biden and Donald Trump won 51.3% and 46.8% of the popular vote respectively (Federal Election Commission, 2022). This dataset slightly over-represents Democrat voters.

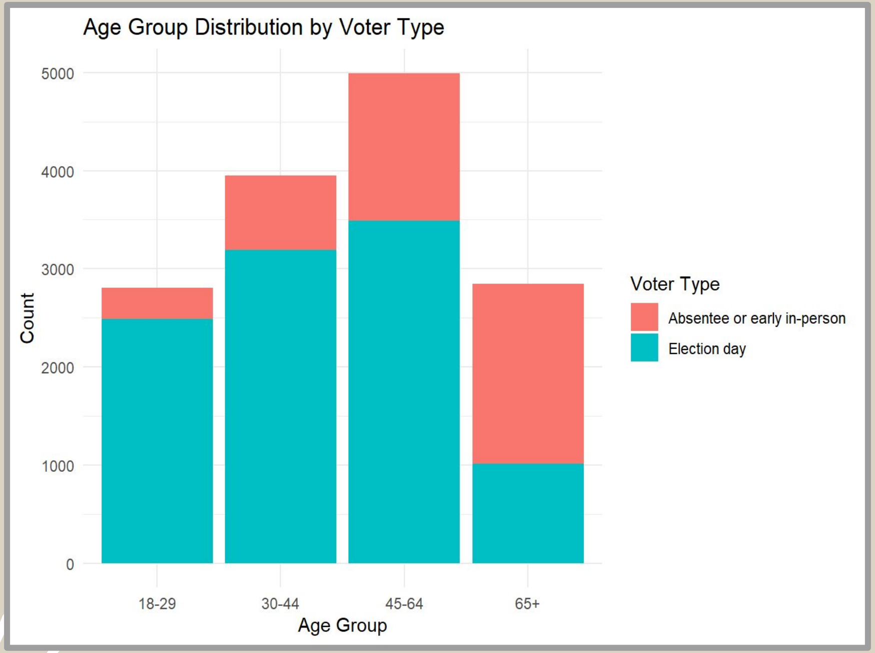
Voter Types

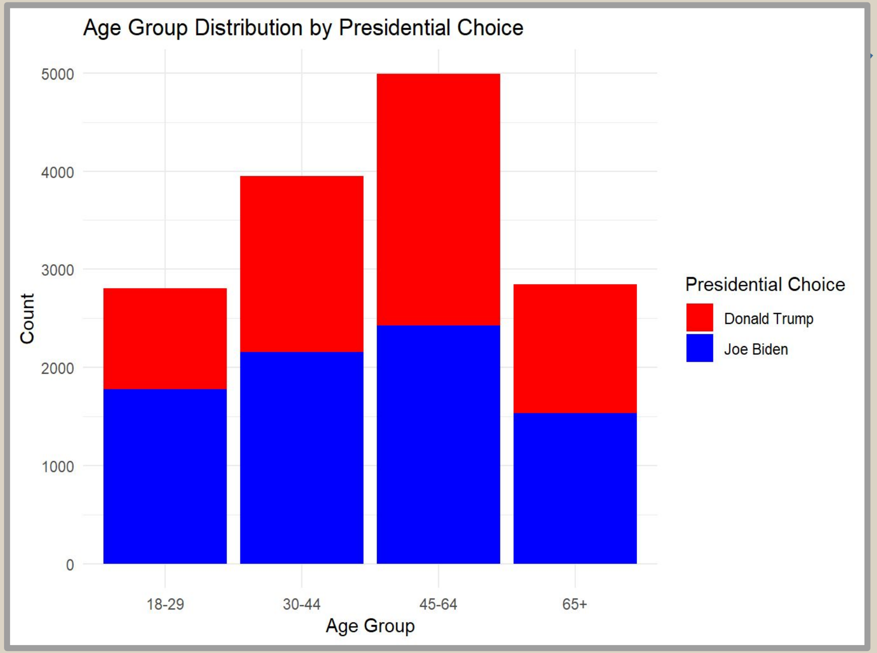




- A majority of voters cast their votes on Election Day.
- Biden leads in for voters who were absent or chose early voting, while both candidates show strong Election Day participation.

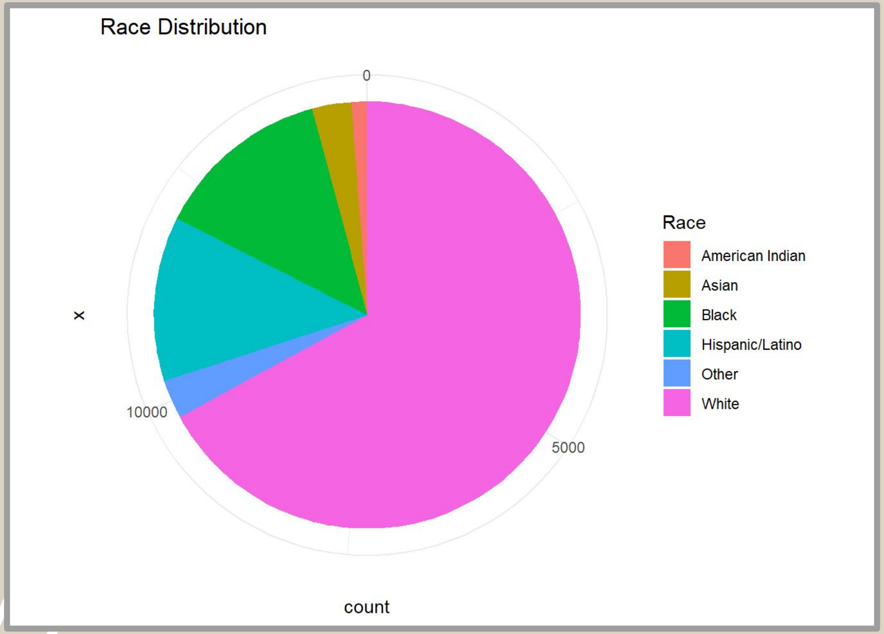
Age Distribution of Voter Types

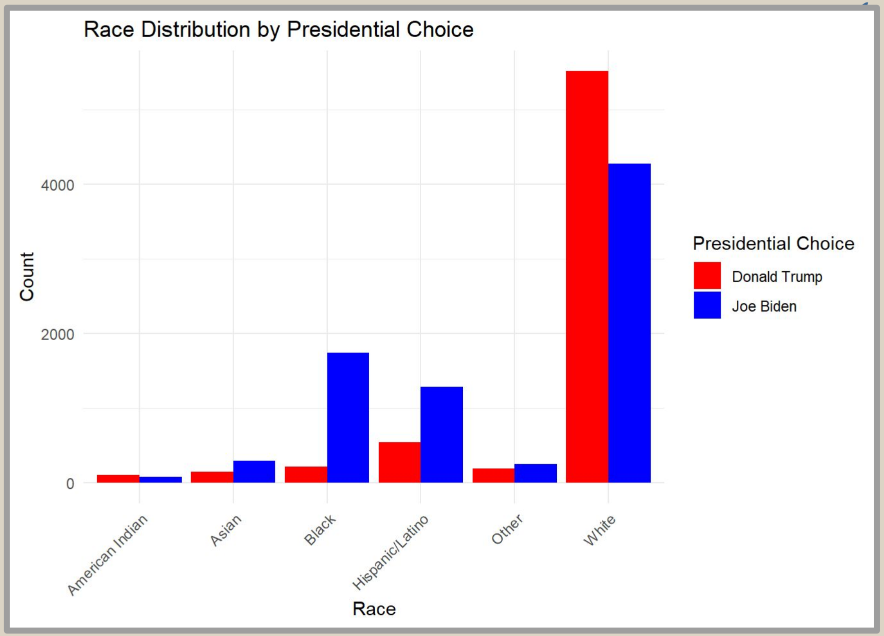




- Election Day voting is the preferred method across all age groups.
- Biden leads among younger voters, while Trump dominates in the 45-64 age group.

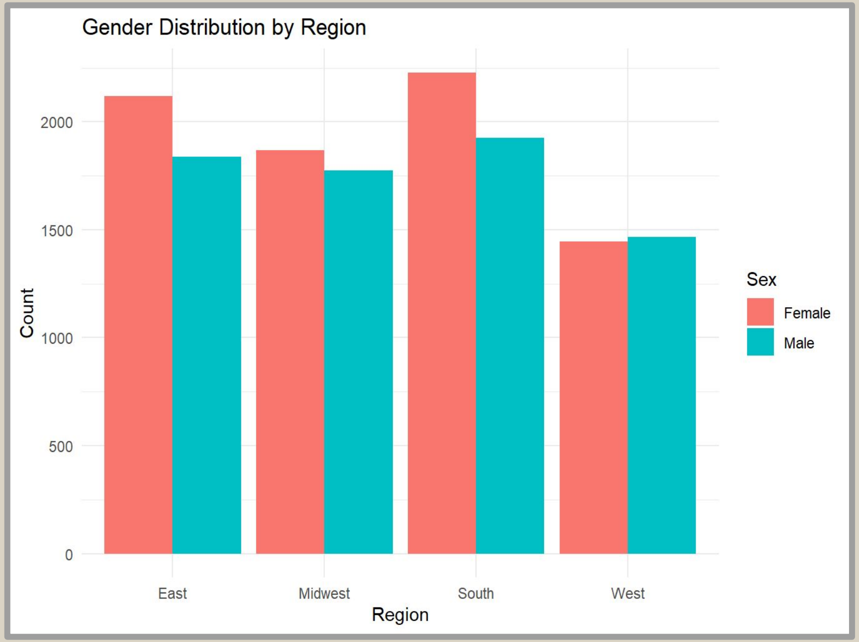
Race Distribution of Voter Types

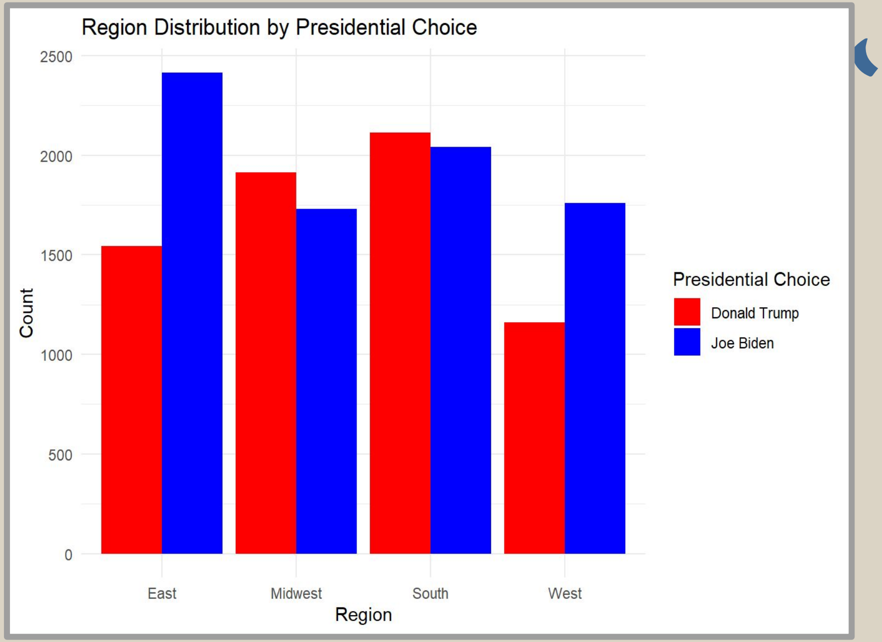




- Biden dominates among Black, Hispanic/Latino, and Asian voters, while Trump holds a strong lead among White voters.
- White voters make up a majority of the voting population.

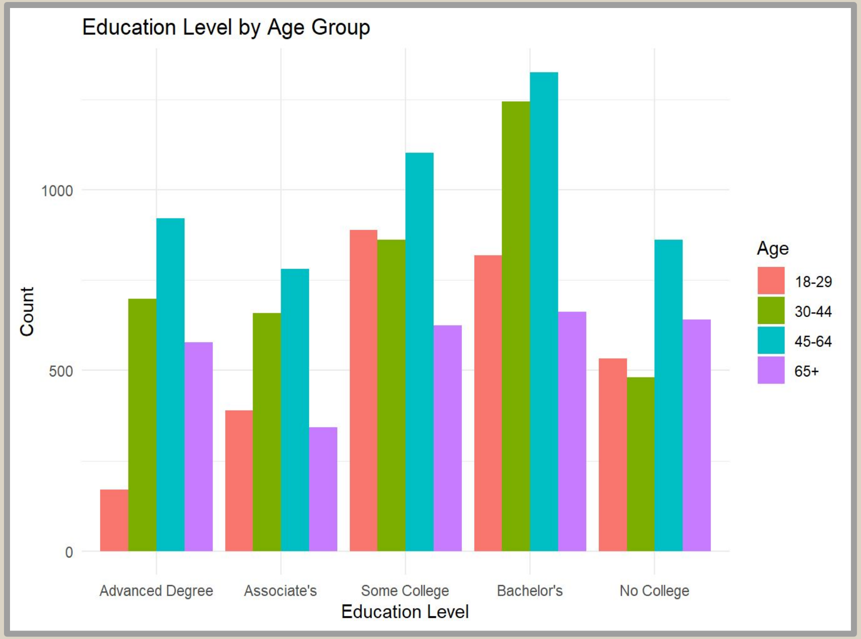
Gender Distribution by Region

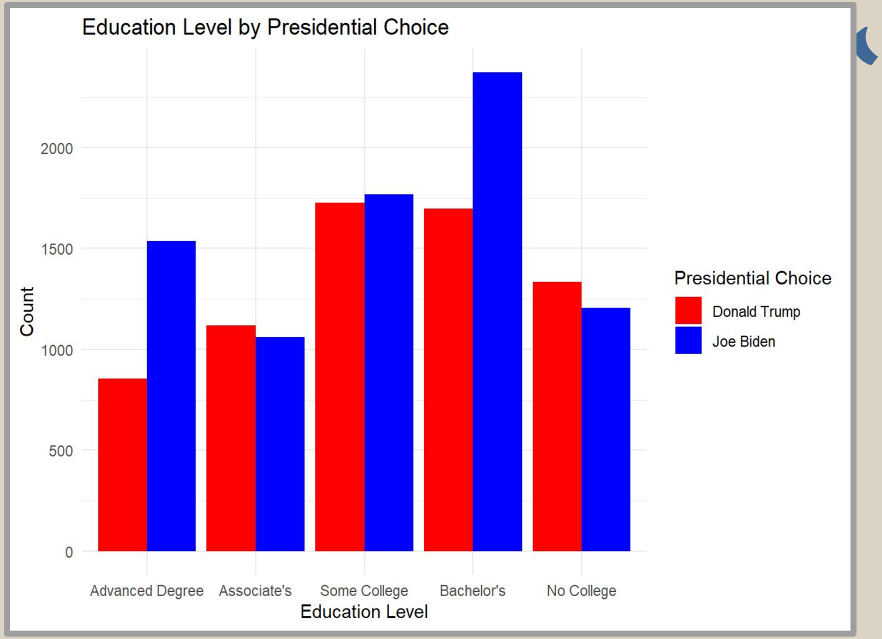




- Gender distribution is relatively balanced across regions
- Biden leads in the East and West regions, while Trump shows stronger support in the South, with the Midwest being nearly evenly split between the two candidates.

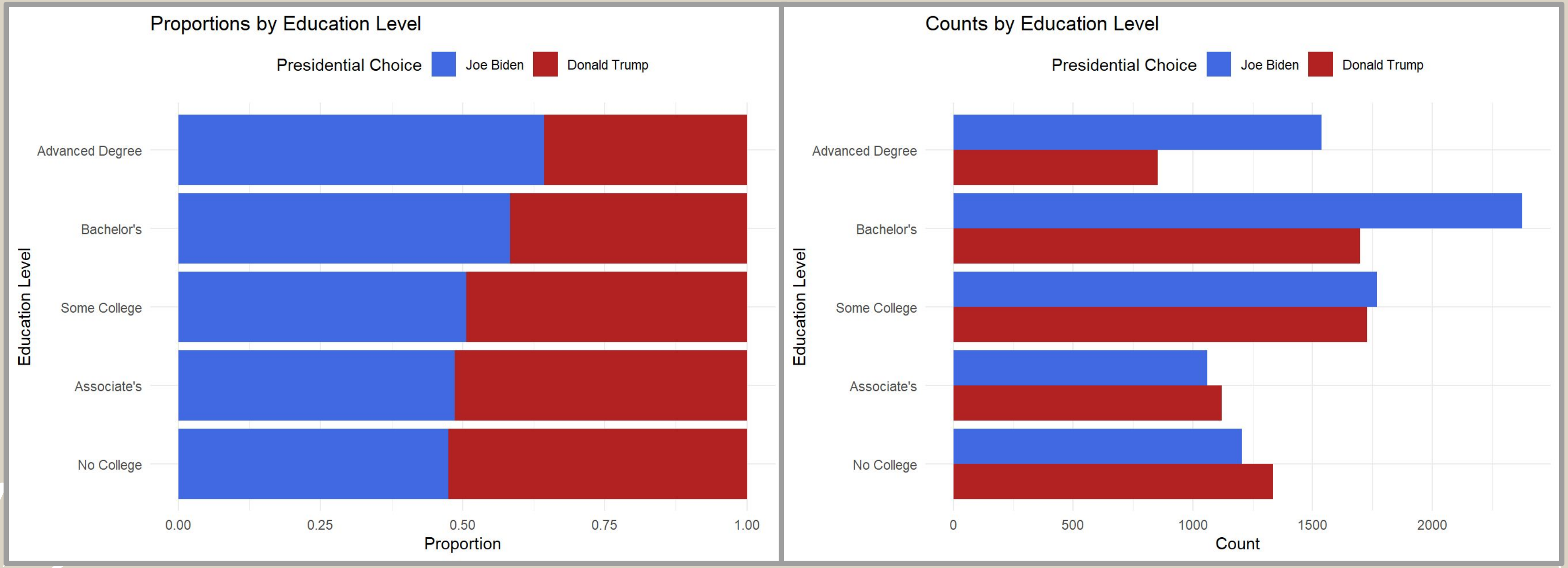
Education Level by Age Group





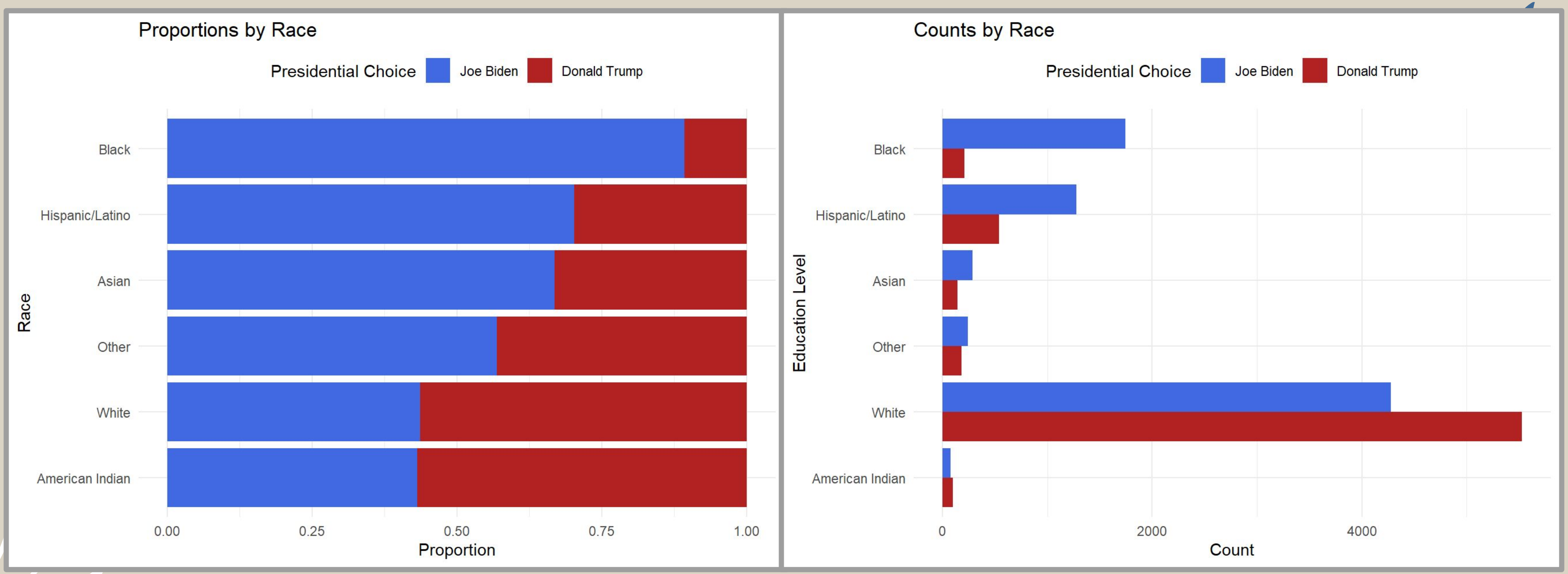
- Bachelors are the most common educational attainment across age groups
- Biden leads among voters with Advanced and Bachelor's degrees, while Trump shows stronger support among those with No College and Associate’s.

Education Level and Presidential Choice



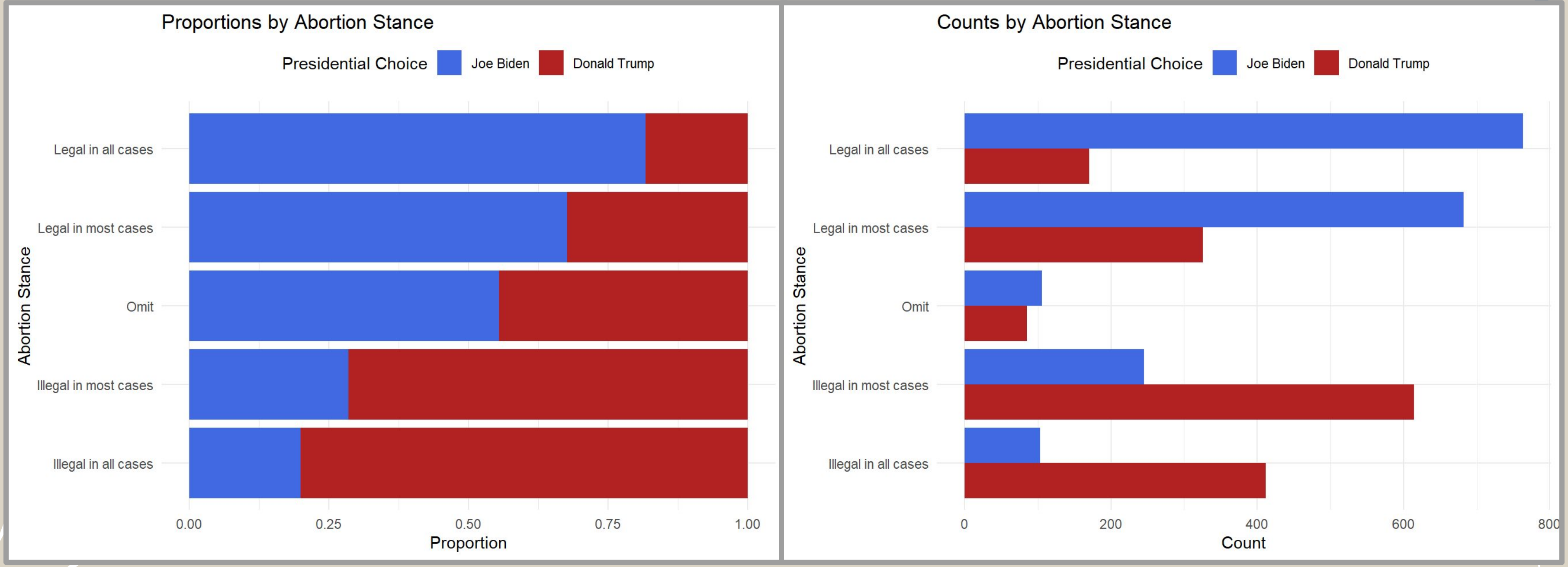
- Biden performs better among voters with Advanced and Bachelor's degrees, while Trump leads among those with No College education or Some College experience.

Race and Presidential Choice



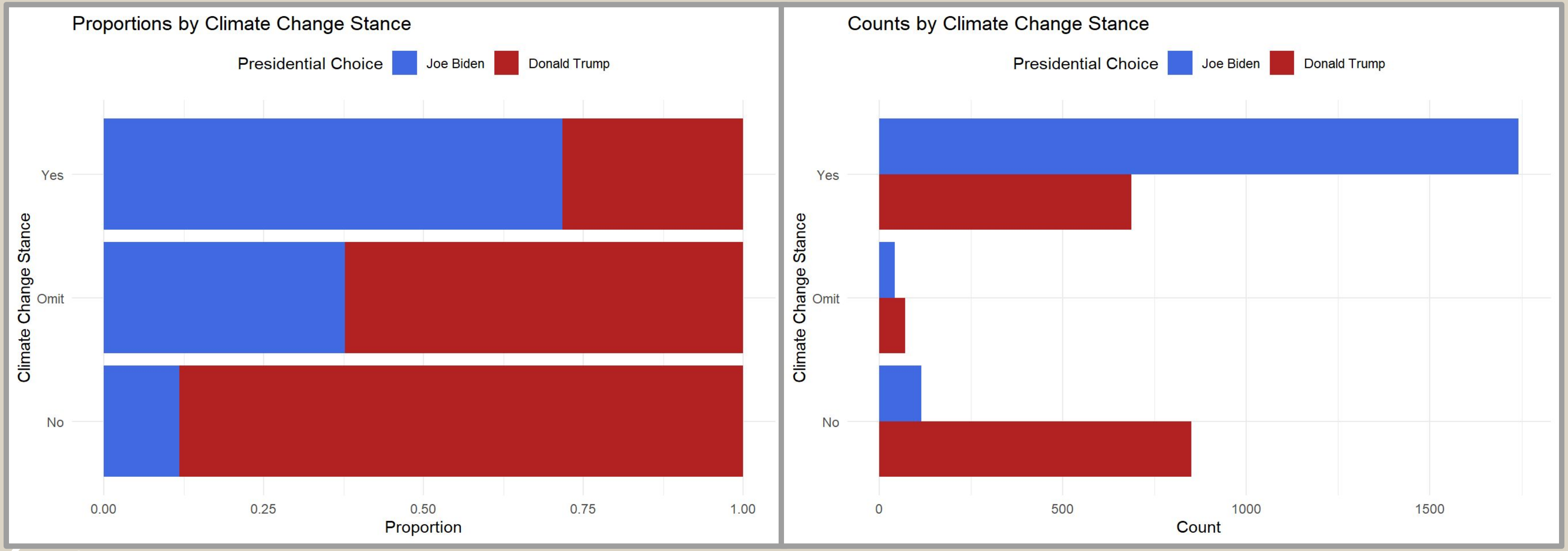
- Biden dominates among Black, Hispanic/Latino, and Asian voters, while Trump holds a strong lead among White voters.
- White voters make up a majority of the voting population.

Abortion Stance and Presidential Choice



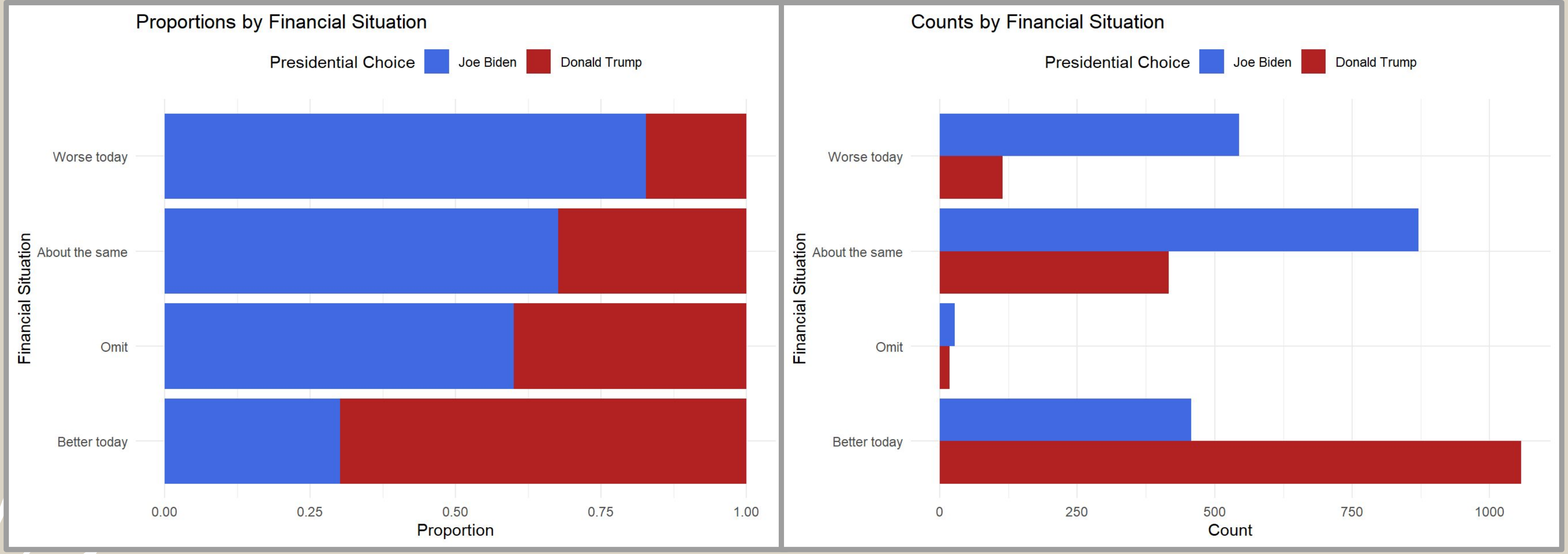
- Biden has more support from voters favoring legal abortion, while Trump dominates among those opposing abortion in most or all cases.

Climate Change Stance and Presidential Choice



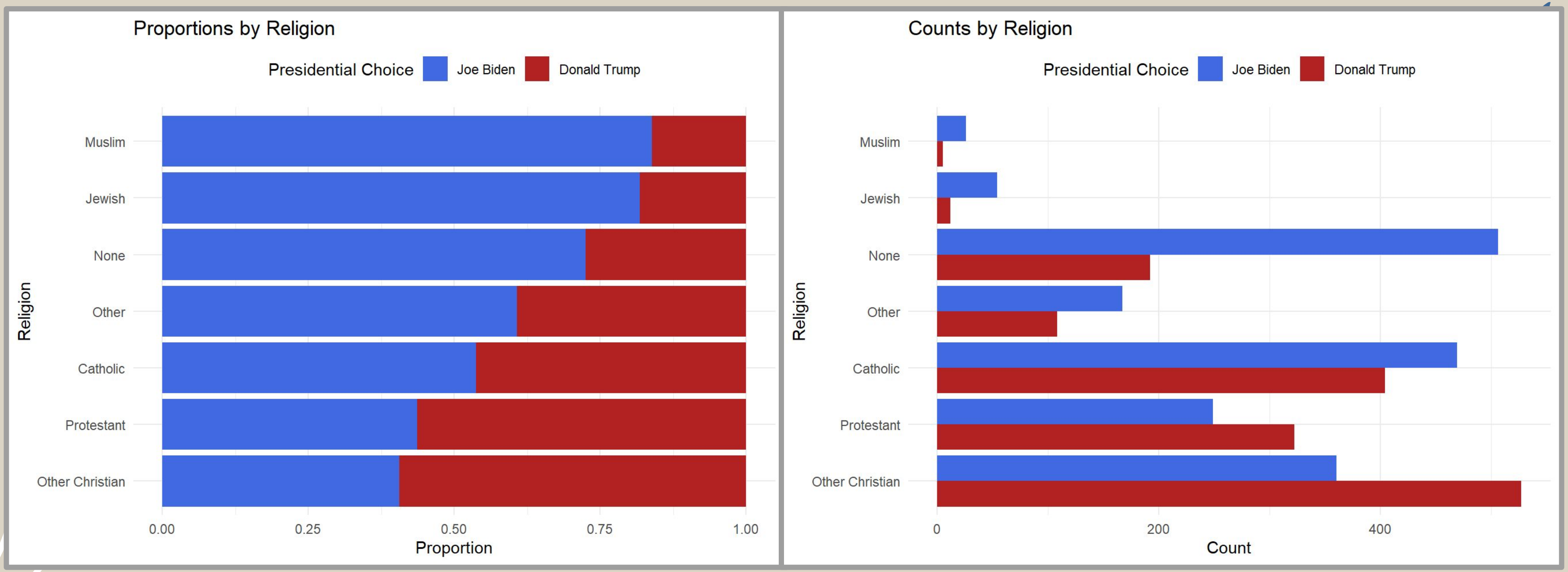
- Biden overwhelmingly leads among voters who believe in climate change, while Trump dominates among voters who do not believe in climate change.
- Very few voters are unwilling to voice a stance on climate change.

Financial Situation and Presidential Choice



- Biden leads among voters who report a worse financial situation, while Trump has more support among those who feel their financial situation is better today.
- Very few voters are unwilling to voice their opinion on their financial situation.

Religion and Presidential Choice



- Biden leads among voters with no religious affiliation and non-Christian religions.
- Trump shows strong support among Christian-related religions.
- Muslim and Jewish voters have low counts and are underrepresented.

3

Model Selection

Candidate Models

Model	Description	Predictors
Personal	11 Personal Demographic Predictors	age, educ18, earlyvel, qraceai, region, sex, sizeplac, lgbt, married, vetvoter
Attitude	4 Social Attitude Predictors	abortion, climatec, finsit, relign18
Combined Full	All 15 Predictors	age, educ18, earlyvel, qraceai, region, sex, sizeplac, lgbt, married, child12, vetvoter, abortion, climatec, finsit, relign18
Combined Reduced	Top 6 Predictors Chosen by Random Forest	educ18, qraceai, abortion, climatec, finsit, relign18

* More models were tested, but are not included as a candidate model due to lack of relevance to the research question.

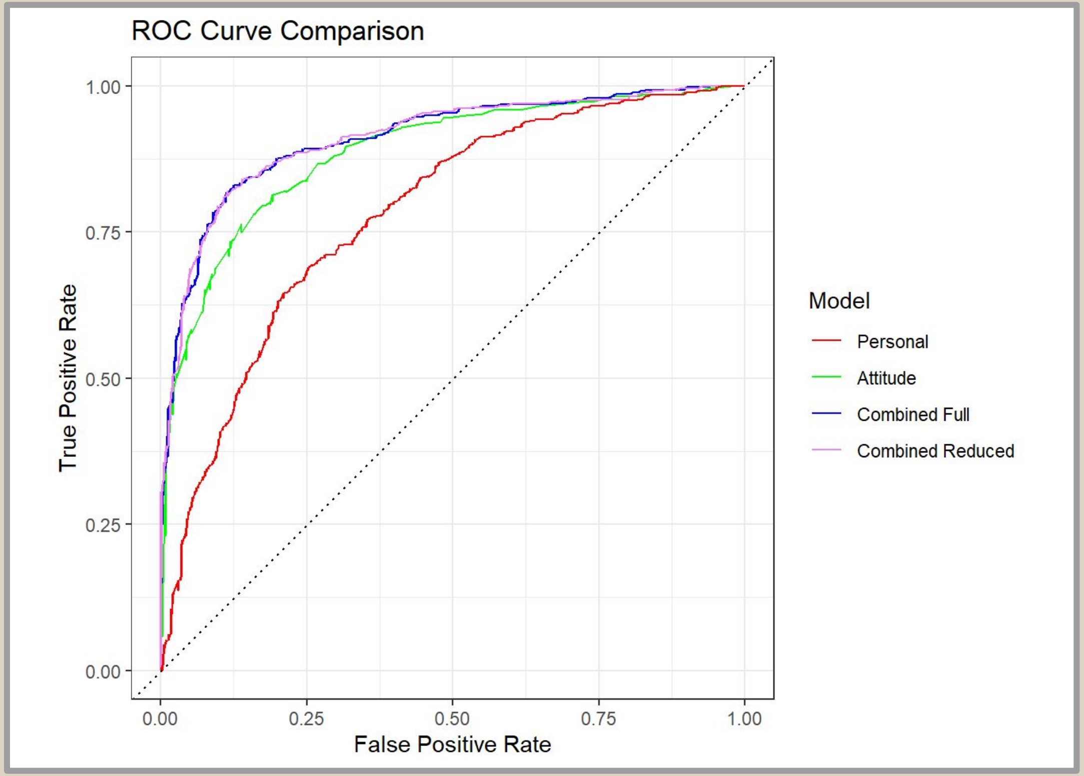
** See appendix for full model comparison.

Model Comparison

Model	Null Deviance	Residual Deviance	Degrees of Freedom	AIC	Accuracy	ROC AUC
Personal	3052.386	2574.831	2188	2624.831	0.7168421	0.7809508
Attitude	3052.386	1871.328	2197	1903.328	0.8126316	0.8862047
Combined Full	3052.386	1654.470	2172	1736.470	0.8536842	0.9107376
Combined Reduced	3052.386	1677.646	2188	1727.646	0.8515789	0.9112934

Data Split: A 0.7 train/test split was used to measure performance.
5-Fold CV: All models used logistic regression with 5-fold cross-validation and selection by accuracy.
Overfitting: Combined Full improves accuracy over Combined Reduced by 0.002 (0.2%) but decreases degrees of freedom by 16 and ROC AUC by 0.0006.
Model Strength: Combined Reduced performs nearly as well as Combined Full with fewer predictors, avoiding overfitting.
Conclusion: Combined Reduced is preferred for its simplicity and strong predictive performance (lowest AIC and highest ROC AUC).

ROC Curves



Combined Full and Combined Reduced have similar ROC curves.

Combined Reduced has the highest ROC AUC value.

Model	ROC AUC
Personal	0.7809508
Attitude	0.8862047
Combined Full	0.9107376
Combined Reduced	0.9112934

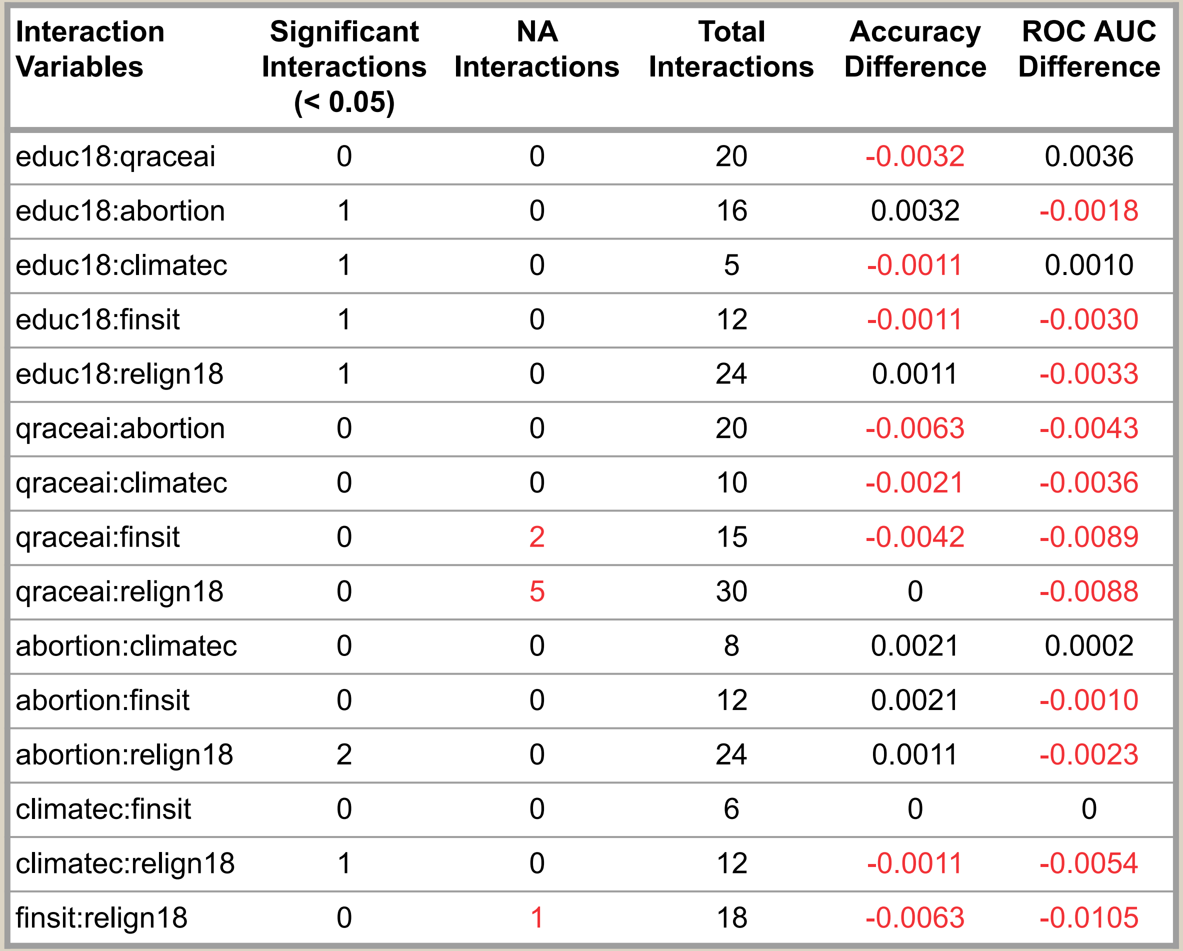
Interaction Effects

Method: All 2-factor interactions were individually tested on Combined Reduced.

Findings: All interactions between 2-variables returned more insignificant predictors than significant predictors. Some interactions were unstable and returned predictors with NAs.

Difference: The maximum increase in accuracy was 0.0032 (0.32%).

Conclusion: Interactions excluded to improve model interpretability with a minimal loss in performance.



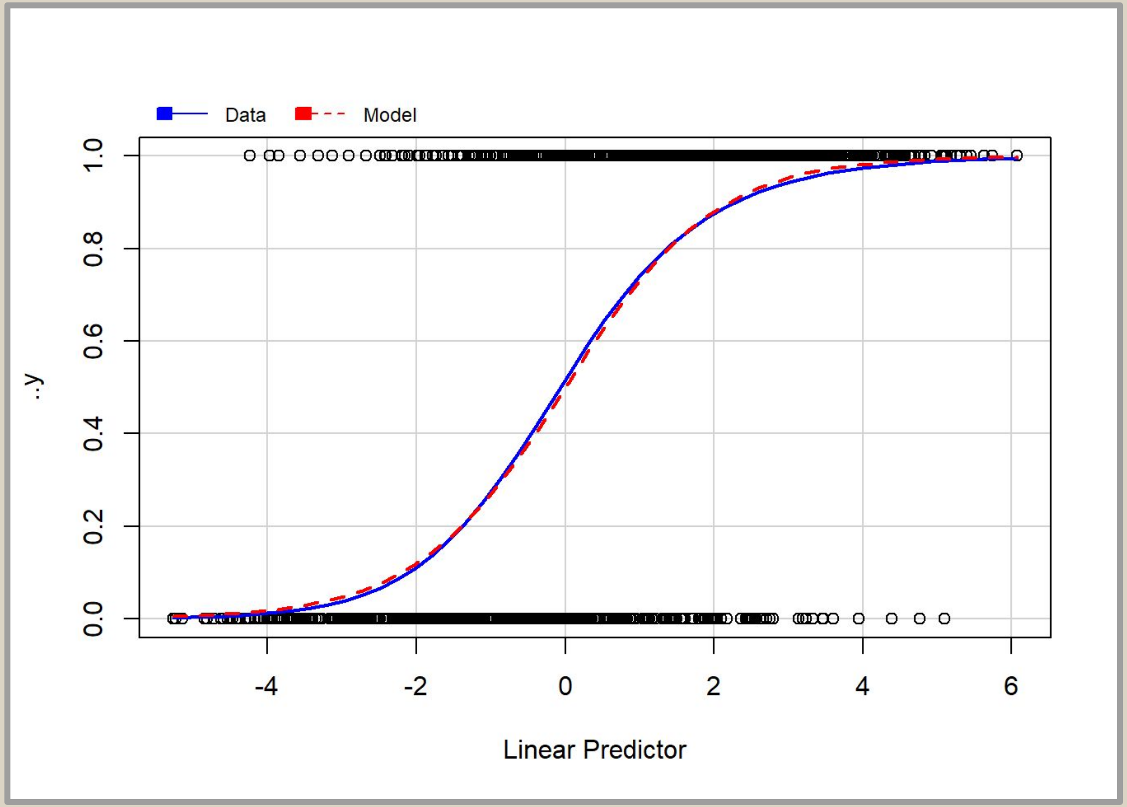
Interaction Variables	Significant Interactions (< 0.05)	NA Interactions	Total Interactions	Accuracy Difference	ROC AUC Difference
educ18:qraceai	0	0	20	-0.0032	0.0036
educ18:abortion	1	0	16	0.0032	-0.0018
educ18:climatec	1	0	5	-0.0011	0.0010
educ18:finsit	1	0	12	-0.0011	-0.0030
educ18:relign18	1	0	24	0.0011	-0.0033
qraceai:abortion	0	0	20	-0.0063	-0.0043
qraceai:climatec	0	0	10	-0.0021	-0.0036
qraceai:finsit	0	2	15	-0.0042	-0.0089
qraceai:relign18	0	5	30	0	-0.0088
abortion:climatec	0	0	8	0.0021	0.0002
abortion:finsit	0	0	12	0.0021	-0.0010
abortion:relign18	2	0	24	0.0011	-0.0023
climatec:finsit	0	0	6	0	0
climatec:relign18	1	0	12	-0.0011	-0.0054
finsit:relign18	0	1	18	-0.0063	-0.0105

4

Model Assumptions

Combined Reduced Model

Marginal Model Plot

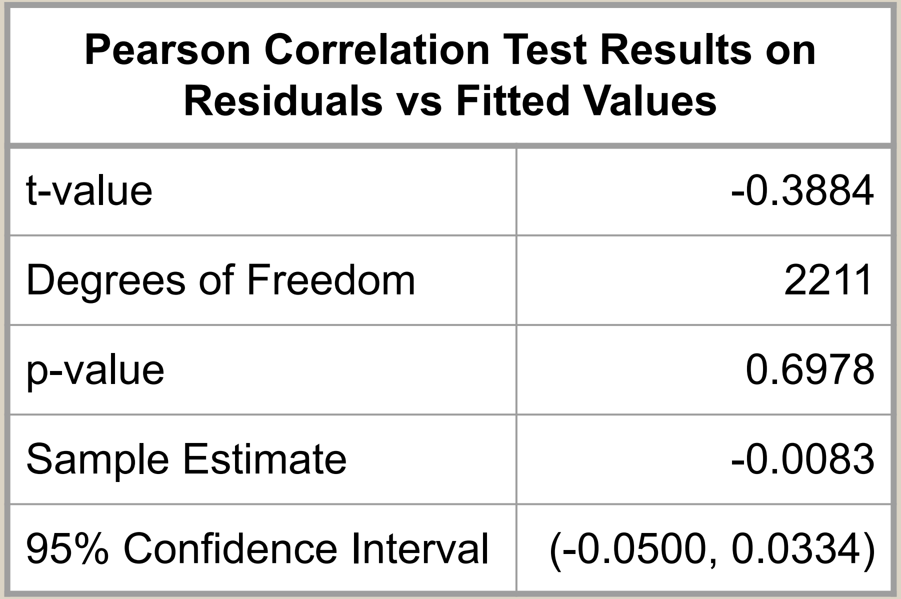


Combined Reduced Model

Purpose: The MMP checks the linearity assumption between predictors and the log-odds in the logistic regression model.

Interpretation: The red dashed line (model fit) closely follows the blue line (data), indicating that the linearity assumption holds for the predictors.

Pearson Correlation Test on Residuals



Pearson Correlation Test Results on Residuals vs Fitted Values

t-value	-0.3884
Degrees of Freedom	2211
p-value	0.6978
Sample Estimate	-0.0083
95% Confidence Interval	(-0.0500, 0.0334)

Combined Reduced Model

Purpose: Assess whether the residuals are randomly distributed.

Findings: Since p-value = 0.6978 > 0.05, we cannot reject the null hypothesis that the true correlation between residuals and fitted values is 0, so there is no significant patterns between the residuals and fitted values.

Conclusion: The model fulfills the assumption that the residuals are randomly distributed.

Generalized Variance Inflation Factors (GVIF)

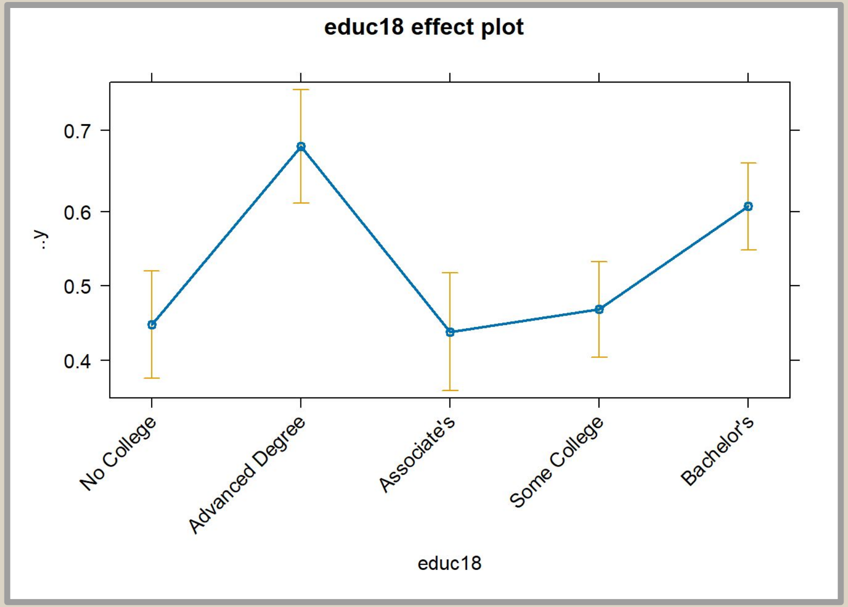
Predictor	GVIF	Degrees of Freedom (DF)	$(GVIF)^{1/(2\times DF)}$
educ18	1.185361	4	1.021483
qraceai	1.401372	5	1.034321
abortion	1.208750	4	1.023981
climatec	1.141615	2	1.033665
finsit	1.106468	3	1.017005
relign18	1.431146	6	1.030324

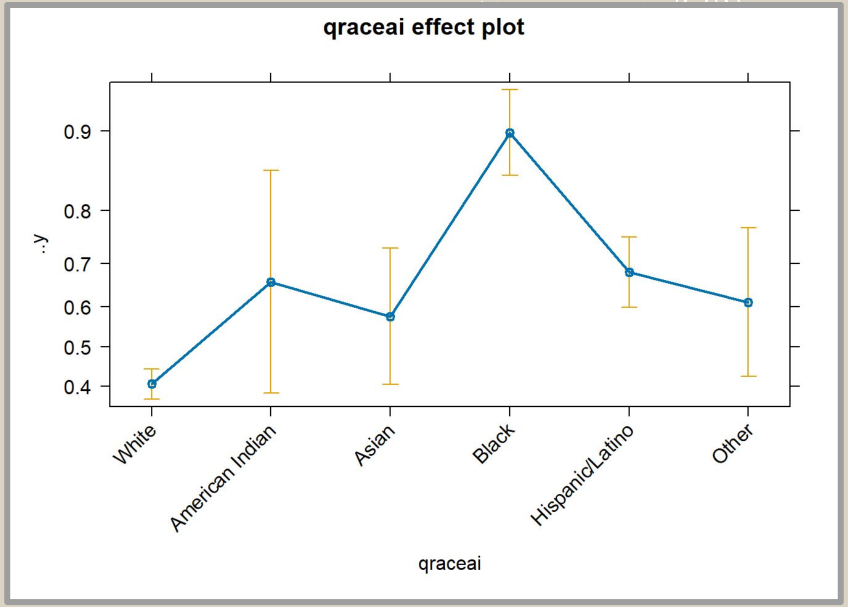
Combined Reduced Model

Purpose: GVIF measures multicollinearity, assessing whether predictors are highly correlated with each other**.**

Interpretation: All GVIF values are below 2, indicating low multicollinearity and no significant issues with predictor redundancy.

Effect Plots – 1

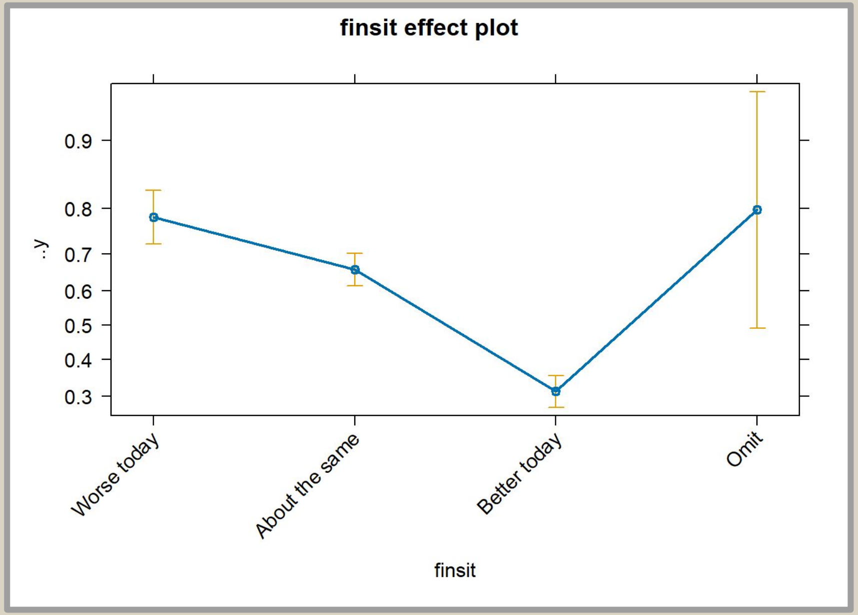


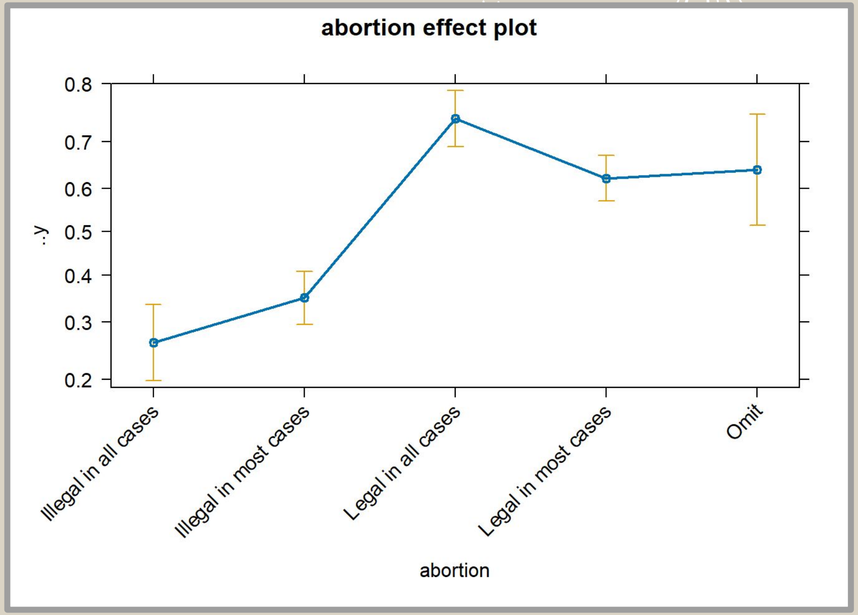


Individuals with an Advanced Degree are the most likely to vote Democrat, followed by those with a Bachelor's Degree and Some College education. Those with Associate's Degrees show a lower likelihood, and individuals with No College education are the least likely to vote Democrat.

Black individuals are the most likely to vote Democrat, followed by American Indian, Hispanic/Latino, and Other racial groups. Asian individuals are less likely, and White individuals are the least likely to vote Democrat.

Effect Plots – 2

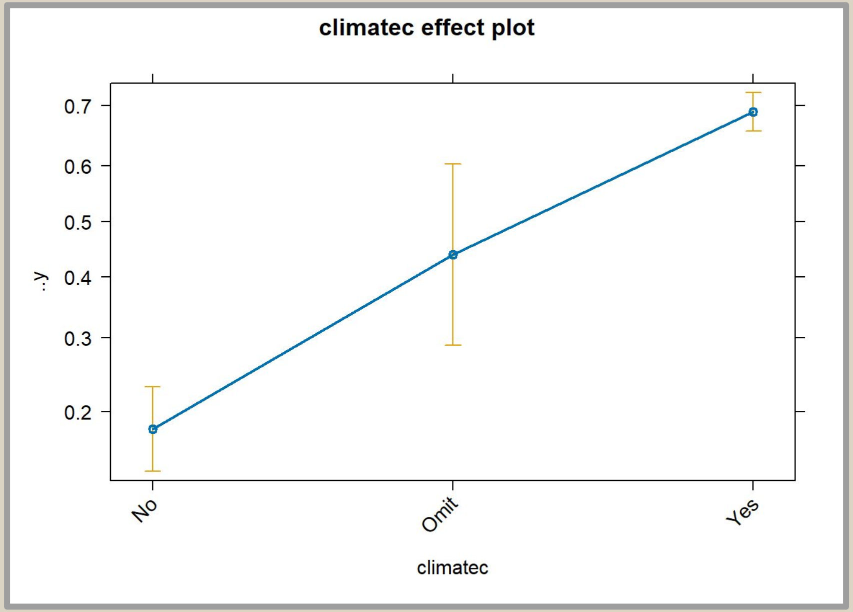




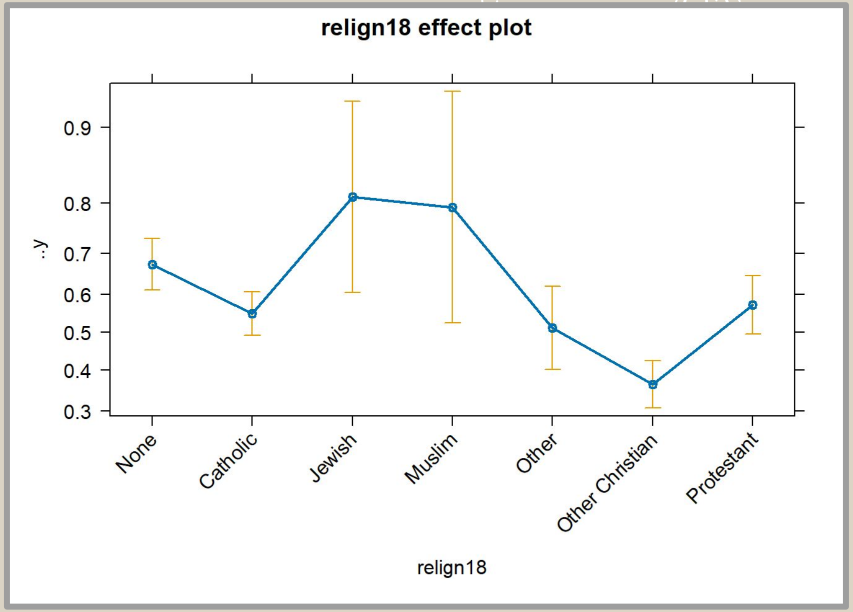
Individuals who feel their financial situation is worse today are the most likely to vote Democrat. Those who believe their financial situation is better today are less likely to vote Democrat.

Individuals who believe abortion should be legal in all cases are the most likely to vote Democrat. Those who think it should be illegal in all cases are the least likely to vote Democrat.

Effect Plots – 3

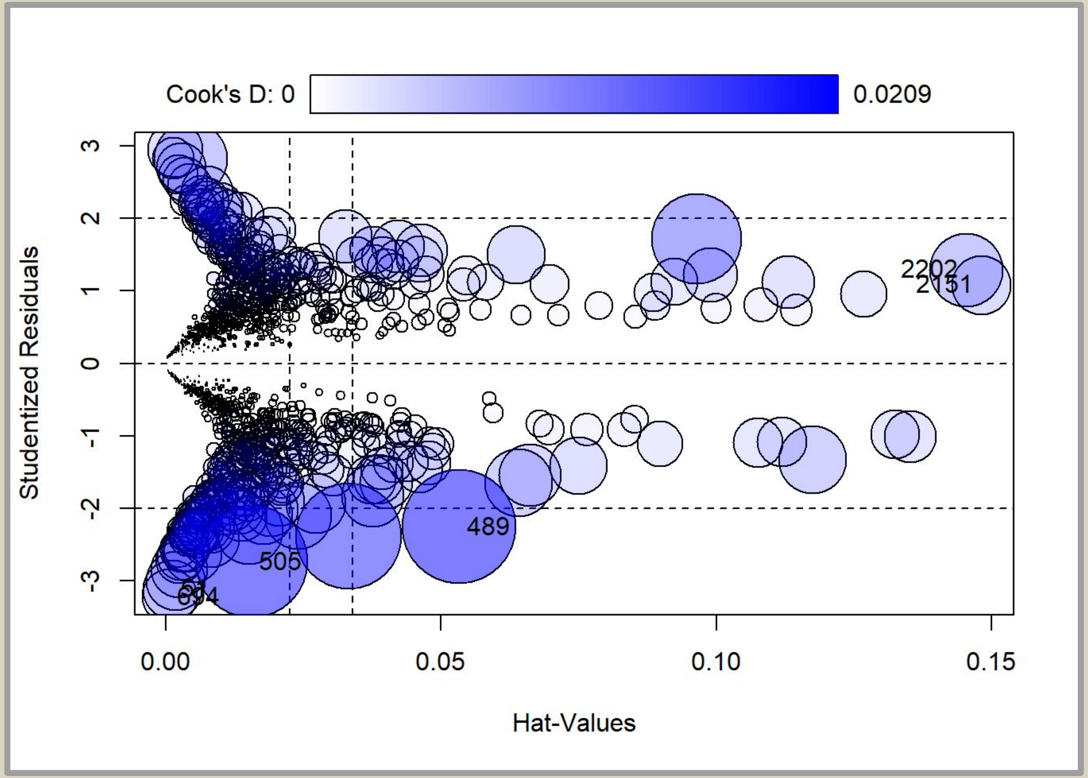


Individuals who believe climate change is a real issue are the most likely to vote Democrat. Those who do not believe in climate change are the least likely to vote Democrat.



Individuals identifying as Jewish or Muslim are the most likely to vote Democrat. Other Christian groups and Other religious affiliations are less likely to vote Democrat.

Influence Plot



Combined Reduced Model

Purpose: Assesses the relationship between leverage (h-values) and residuals for diagnosing model fit.

Findings: There are 2 influential points and 4 outliers. A model was tested without these points, but since the training sample size was large (2,213), there was no improvement to accuracy or ROC AUC.

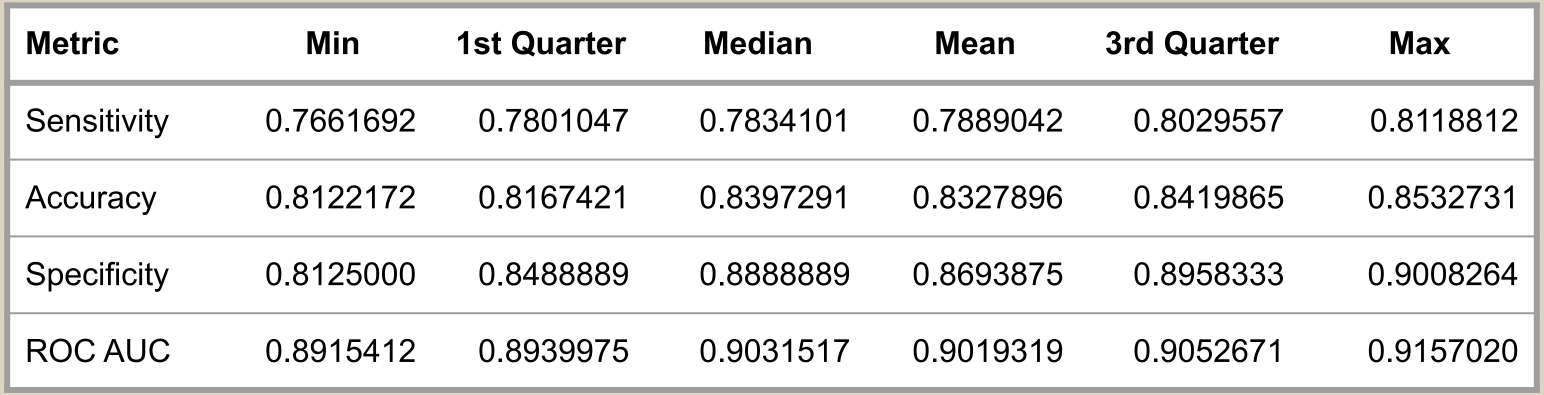
** See appendix for full model comparison.

5

Results

Combined Reduced Model

Cross Validation Statistics for Combined Reduced



Metric	Min	1st Quarter	Median	Mean	3rd Quarter	Max
Sensitivity	0.7661692	0.7801047	0.7834101	0.7889042	0.8029557	0.8118812
Accuracy	0.8122172	0.8167421	0.8397291	0.8327896	0.8419865	0.8532731
Specificity	0.8125000	0.8488889	0.8888889	0.8693875	0.8958333	0.9008264
ROC AUC	0.8915412	0.8939975	0.9031517	0.9019319	0.9052671	0.9157020

Cross-validation helps estimate the model's generalization performance on unseen data by splitting the data into training and validation subsets multiple times.

5-fold cross-validation was deemed most appropriate given that the dataset has around 3,000-4,000 observations per survey version.

Findings: The model performs better at identifying Republican voters (specificity) than Democrat voters (sensitivity).

Evaluation: To minimize bias, accuracy is the preferred selection metric.

Odds Ratios Plot

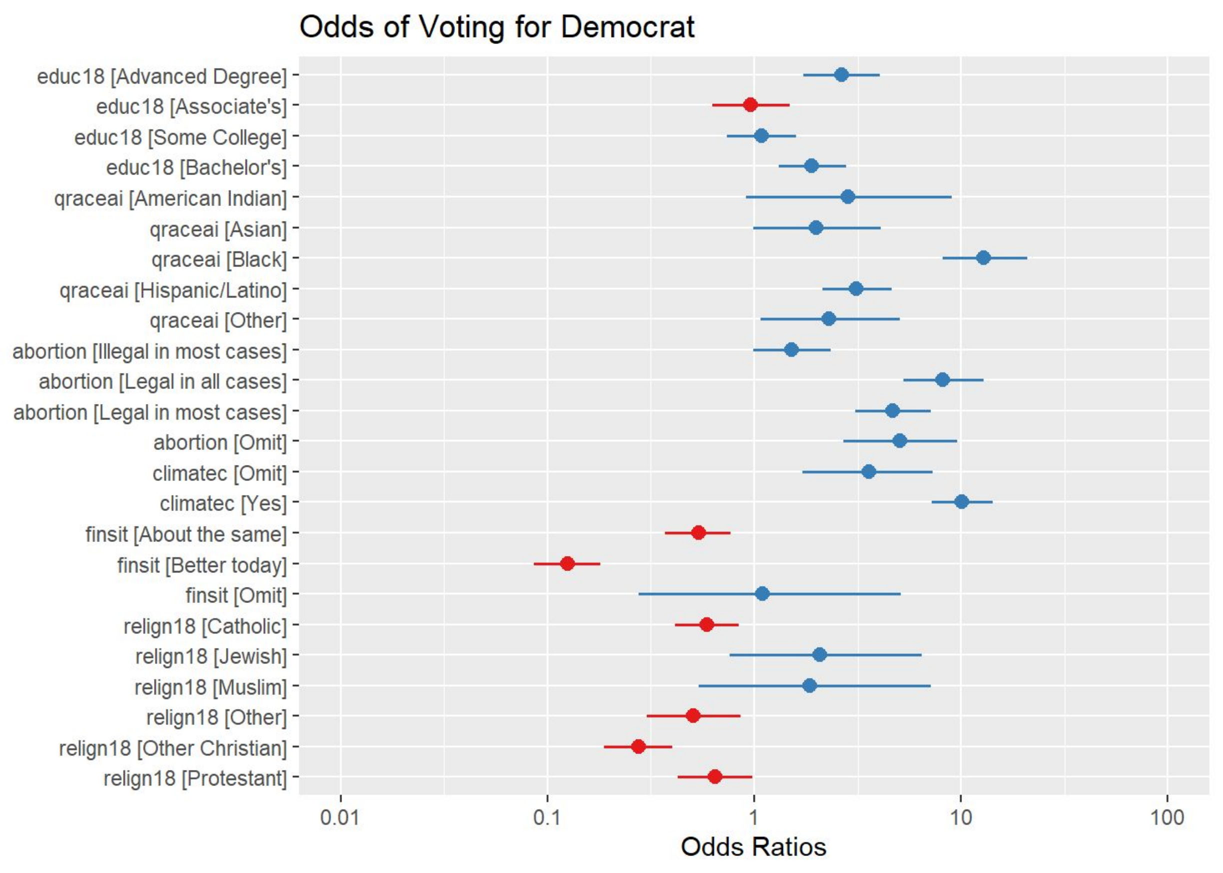
Combined Reduced Model

Reference Levels:
- educ18: **No College**
- qraceai: **White**
- abortion: **Illegal in all cases**
- climatec: **No**
- finsit: **Worse today**
- relign18: **None**

Plot Interpretation

BLUE:
- **Increased** odds of voting Democrat. Further right indicates higher odds.

RED:
- **Decreased** odds of voting Democrat. Further left indicates lower odds.



Odds Ratio Plot Findings

Education

- With the exception of Associate degree holders, higher education levels are associated with increased odds of voting for Democrats.

Race

- Voters who are People of Color are more likely to to vote for Democrats than White voters.

Abortion Stance

- Stronger stances towards abortion being legal are associated with increased odds of voting for Democrats.
- Voters who chose to omit their stance on abortion are more likely to vote Democrats than voters who believe abortion should be illegal in most cases.

Climate Change Stance

- Voters who believe in climate change or omitted their stance on climate change are more likely to vote for Democrats than those voters who do not believe in climate change.

Financial Situation

- Increased positive sentiments about a voter's financial situation decreases their odds of voting for Democrats.
- Voters who chose to omit their financial situation information have about the same odds of voting for Democrats as voters who feel negatively about their financial situation

Religion

- Jewish and Muslim voters are more likely to vote for Democrats than voters with no religion.
- Voters who are Christian and other religions are less likely to vote for Democrats than voters with no religion.

Odds Ratios Table

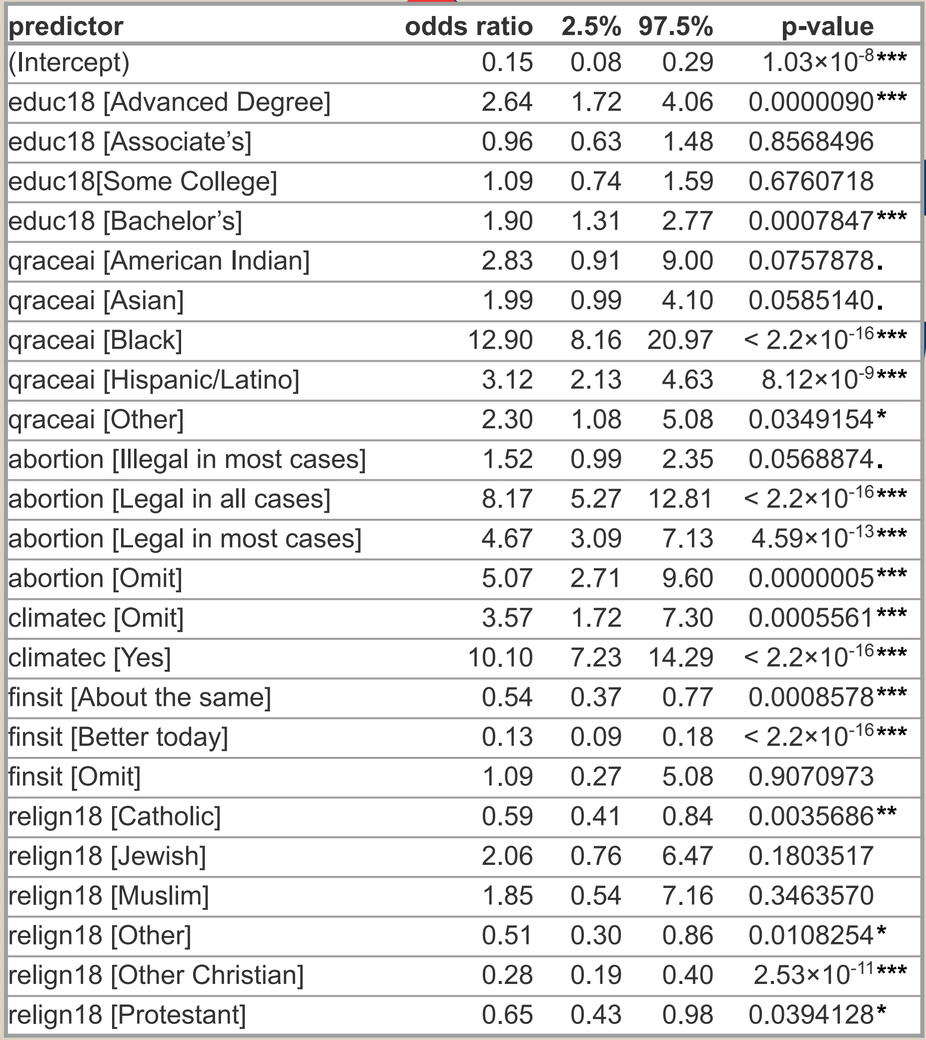
Combined Reduced Model

Reference Levels:

- educ18: **No College**
- qraceai: **White**
- abortion: **Illegal in all cases**
- climatec: **No**
- finsit: **Worse today**
- relign18: **None**

Interpretation Example:

On average, “Advanced Degree” holders are 2.6 times more likely to vote for Democrats than voters with “No College.” We are 95% confident that the odds of an “Advanced Degree” holder voting for Democrats is between 1.7–4 times more than voters with “No College.” These odds are extremely significant.



predictor	odds ratio	2.5%	97.5%	p-value
(Intercept)	0.15	0.08	0.29	1.03×10^{-8}***
educ18 [Advanced Degree]	2.64	1.72	4.06	0.0000090***
educ18 [Associate’s]	0.96	0.63	1.48	0.8568496
educ18[Some College]	1.09	0.74	1.59	0.6760718
educ18 [Bachelor’s]	1.90	1.31	2.77	0.0007847***
qraceai [American Indian]	2.83	0.91	9.00	0.0757878.
qraceai [Asian]	1.99	0.99	4.10	0.0585140.
qraceai [Black]	12.90	8.16	20.97	$< 2.2\times10^{-16}$***
qraceai [Hispanic/Latino]	3.12	2.13	4.63	8.12×10^{-9}***
qraceai [Other]	2.30	1.08	5.08	0.0349154*
abortion [Illegal in most cases]	1.52	0.99	2.35	0.0568874.
abortion [Legal in all cases]	8.17	5.27	12.81	$< 2.2\times10^{-16}$***
abortion [Legal in most cases]	4.67	3.09	7.13	4.59×10^{-13}***
abortion [Omit]	5.07	2.71	9.60	0.0000005***
climatec [Omit]	3.57	1.72	7.30	0.0005561***
climatec [Yes]	10.10	7.23	14.29	$< 2.2\times10^{-16}$***
finsit [About the same]	0.54	0.37	0.77	0.0008578***
finsit [Better today]	0.13	0.09	0.18	$< 2.2\times10^{-16}$***
finsit [Omit]	1.09	0.27	5.08	0.9070973
relign18 [Catholic]	0.59	0.41	0.84	0.0035686**
relign18 [Jewish]	2.06	0.76	6.47	0.1803517
relign18 [Muslim]	1.85	0.54	7.16	0.3463570
relign18 [Other]	0.51	0.30	0.86	0.0108254*
relign18 [Other Christian]	0.28	0.19	0.40	2.53×10^{-11}***
relign18 [Protestant]	0.65	0.43	0.98	0.0394128*

Odds Ratio Table Findings

Education

- Individuals with advanced degrees are 2.64 times more likely to vote Democrat compared to those without college education. More specifically, voters with a bachelor’s degree are 1.90 times more likely to vote Democrat. This is a clear educational divide.

Race

- Black voters are 12.90 times more likely to vote Democrat compared to White individuals and Hispanic/Latino voters are 3.12 times more likely to vote Democrat. This highlights a remarkable disparity that race is a strong predictor of voting behavior.

Abortion Stance

- Voters who support abortion being legal in all cases are a striking 8.17 times more likely to vote Democrat, a sharp partisan divide on this social issue. This may be due to to voters being Republican following a more conservative religion and ideology.

Climate Change Stance

- Despite very few voters being unwilling to voice a stance on climate change, voters who chose to omit their stance on climate change are 3.6 times more likely to vote for Democrats than voters who do not believe in climate change, and these odds are extremely significant.

Financial Situation

- The odds for “Omit” is not statistically significant. This may be due to very few voters being unwilling to voice their opinion on their financial situation.

Religion

- Jewish and Muslim religions are about twice as likely to vote for Democrats than voters with no religion, but these odds are not statistically significant. This may be due to Jewish and Muslim voters being underrepresented in the dataset.

6

Conclusion

Key Findings
Shortcomings
Recommendations

Key Findings

- **Objective**: uncovered trends and patterns in voting behavior through voters' demographic and social attitudes.

- **Influential Variables:** variables associated with education level, race, and social attitudes on topics like abortion are identified as strong predictors of voting decisions.

- **Methodology:** utilized methods like exploratory data analysis, random forest, and logistic regression to conduct voting behavior analysis.

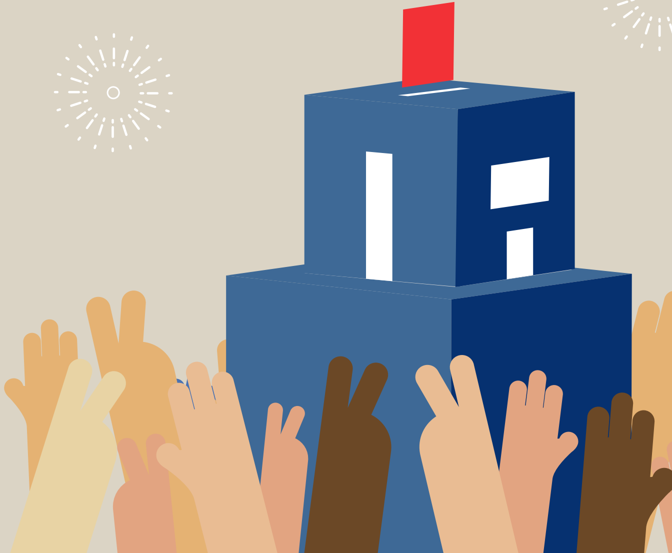
- **Evaluation:** identified the Combined Reduced model to be the best performing model with its low AIC, high prediction accuracy, and high ROC AUC score.

Shortcomings

- **Imbalanced Data**: Limited representations of certain demographic groups may introduce bias into the model predictions.

- **Model Overfitting:** models face potential challenges with overfitting, indicating reduced generalizability in prediction results.

- **Survey Limitations**: the various survey versions and designs may introduce potential biases into survey responses and affect the reliability of the data.



Recommendations

- **Imbalance Adjustment:** consider using weighted models to adjust for the data imbalances.
- **Model Fine Tuning:** look deeper into interaction effects or adopt other methods of feature engineering to enhance predictive performance.
- **Standardized Survey Questions:** minimizing the difference in survey questions and versions to increase effective sample size and representation.

THANK YOU VERY MUCH!

Appendix

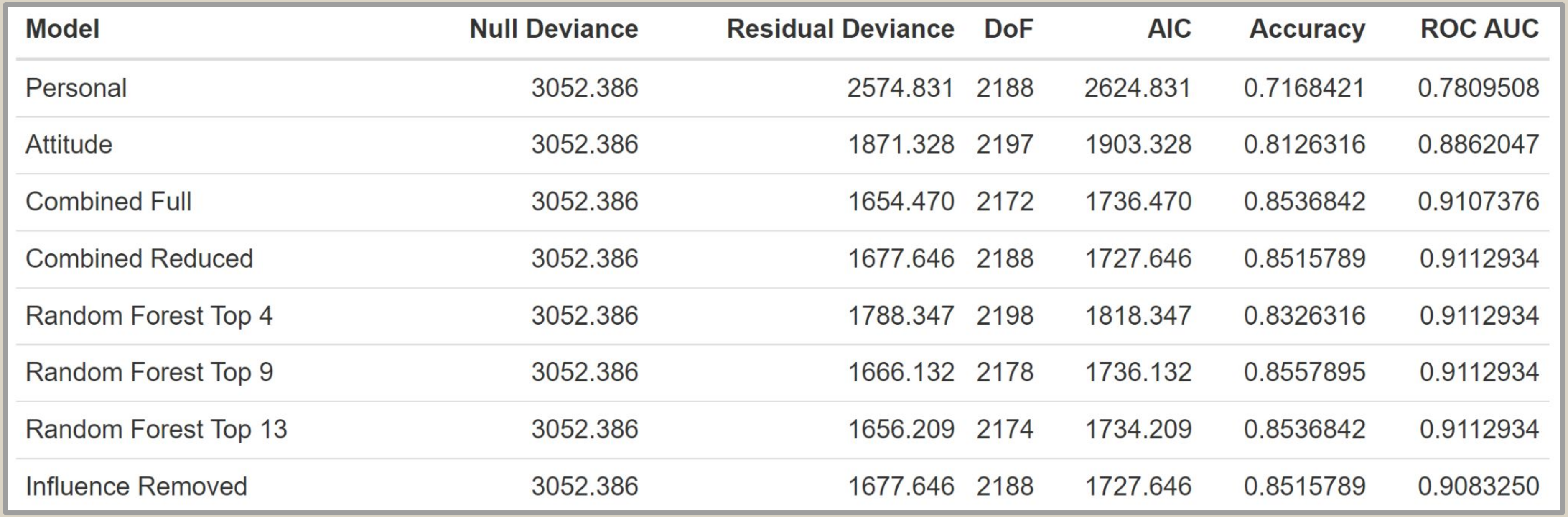
- Citations
- Full Model Comparison

Citations

National Election Pool (ABC News, CBS, CNN, NBC). (2020). National Election Pool Poll: 2020 National Election Day Exit Poll (Version 4) [Dataset]. Roper Center for Public Opinion Research. doi:10.25940/ROPER-31119913

Federal Election Commission. (2022). *Federal Elections 2020: Election Results for the U.S. President, the U.S. Senate and the U.S. House of Representatives*. Federal Election Commission.

Full Model Comparison



Model	Null Deviance	Residual Deviance	DoF	AIC	Accuracy	ROC AUC
Personal	3052.386	2574.831	2188	2624.831	0.7168421	0.7809508
Attitude	3052.386	1871.328	2197	1903.328	0.8126316	0.8862047
Combined Full	3052.386	1654.470	2172	1736.470	0.8536842	0.9107376
Combined Reduced	3052.386	1677.646	2188	1727.646	0.8515789	0.9112934
Random Forest Top 4	3052.386	1788.347	2198	1818.347	0.8326316	0.9112934
Random Forest Top 9	3052.386	1666.132	2178	1736.132	0.8557895	0.9112934
Random Forest Top 13	3052.386	1656.209	2174	1734.209	0.8536842	0.9112934
Influence Removed	3052.386	1677.646	2188	1727.646	0.8515789	0.9083250