Exploring Diverse Influences on Voting Patterns: A Study of Trump and Biden Supporters in the 2020 US Presidential Election

Rory O Sullivan
20232721
Department of Mathematics and Statistics
University of Limerick

Supervisor
Dr. XXXX

Report for a Final Year Project
BSc in Mathematical Sciences
December 12, 2024

Contents

Acknowledgements

I would like to acknowledge my supervisor Dr. XXXX for his continued guidance and support during this project. Additionally, I appreciate the collaborative discussions with my group members on this project. Lastly, I extend my thanks to the members of the Statistics Faculty who were involved in the FYP module.

Abstract

This paper explores key demographic, socioeconomic, and political opinion variables shaping voting patterns for Trump and Biden in the 2020 US Presidential Election, drawing insights from influential studies by Sheng, Garand, Abramowitz and more. The analysis of the ANES 2020 Time Series dataset, comprising 5729 entries, reveals intriguing patterns. While age lacks a clear voting preference difference, sex and ethnicity play significant roles, with males favouring Trump. White Americans display diverse preferences, while Black Americans overwhelmingly favour Biden. Socioeconomically, social class exhibits a balanced preference, and Liberal identity shows a strong preference for Biden. Marital status and religious affiliations display interesting patterns, and political opinion variables including abortion, the death penalty, defence spending and immigration policy provide further insights. Employing a Multiple Logistic Regression and Random Forest model, the study highlights significant predictors of candidate choice, notably on opinion variables such as immigration policy, abortion and defence spending. Both models exhibit good performance with the Logistic model seeming a better predictor of Biden votes.

Chapter 1

Introduction

The complex interactions between demographic, socioeconomic, political opinions and voting behaviour have long been a subject of profound interest in political science research. In the context of the 2020 US Presidential Election, with its notable turnout and political divisions, we have an opportunity to examine how individual traits influence voting. This paper embarks on a comprehensive exploration, seeking to identify different demographic, socioeconomic and political opinions that shaped the voting patterns of supporters of the two major candidates, Donald Trump and Joe Biden.

By concentrating on the 2020 election, this research seeks to enhance the existing literature on political behaviour and electoral studies, shedding light on the nuanced dynamics that motivated Trump and Biden supporters to vote.

In navigating this exploration, the paper will draw upon a synthesis of existing literature, employing a critical review of relevant studies to establish a foundation for our analysis. Through an in-depth examination of prior research, this study aims to contextualise its findings within the broader academic discourse, contributing to our collective understanding of the intricate relationships between demographic, socioeconomic, political opinion factors and voting patterns.

Chapter 2

Literature Review

2.1 Group Literature Review

2.1.1 History of the ANES (1)

The American National election studies (ANES) was founded in 1948. Since then, the survey has become a pivotal part in advancing the understanding of electoral behavior and political attitudes in the United States. Each election year an ANES study is carried out, there are two time points. One survey is conducted before the election and the second after the election. This gives us both pre-election survey data and post-election survey data. The ANES is an extremely rich dataset which has been paramount in moulding the landscape of American political science. ANES studies provide us with crucial insights into public opinion and also provide scholars and researchers with valuable insights into the dynamics of American democracy. Over the years the ANES has contributed significantly to academic research such as The American Voter (Campbell et al., 1980), which many believe revolutionized the study of political behaviour [1]. The dataset has been essential in studies done relating to political, voter turnout, and the evolving demographics of American voters. The ANES has remained relevant throughout the year and the survey is built on each year with new questions being added each year. The ANES has shaped the academic discourse on politics (Leighley, 2013) [2]. The ANES is so highly regarded due to a number of principles they follow when constructing survey's that will be discussed below.

2.1.2 Principles of the ANES (1)

The ANES are a highly regarded and trusted entity. There are several key principles that they ensure to follow when carrying out their surveys. ANES is committed to transparency and scholarly rigour, ensuring that it's data collection and processing procedures are well-documented and openly accessible to researchers (Lupia, 2008) [3]. They also put a massive emphasis on the importance of survey design aiming to compile questionnaires that are balanced, unbiased, and non-partisan, to accurately reflect public opinion (Kellstedt et al., 2017) [4]. They prioritize the longitudinal aspect of its studies, enabling scholars to trace changes in political attitudes and behaviors over time (Campbell et al., 1980) [1]. Together, these principles contribute to the ANES' reputation as a trusted and valuable resource for the study of American politics. How exactly the ANES do this will be discussed in the ANES methodologies.

2.1.3 ANES Methodologies (1)

The focus of ANES efforts over the decades has been on Time series studies. Surveys are conducted shortly before and shortly after the American Presidential election. The surveys are carried out by face to face interviews usually however due to the COVID-19 pandemic and the cost of carrying out these interviews,there has been effort to try move these interviews online via phone and web calls.ANES studies use probability samples, in which each person in the target population has a known, nonzero probability of being selected for the study. All ANES studies use procedures known collectively as complex sampling to distinguish them from simple random sampling. (DeBell , 2010) [5]. The ANES use both oversampling and stratified cluster sampling for the following reasons,

Oversampling involves deliberately selecting more individuals from specific subpopulations or groups of interest in a survey sample. These groups are sampled at a higher rate than their proportion in the population (Groves, 2006) [6]. Oversampling ensures that there are sufficient responses from historically underrepresented groups, such as racial and ethnic minorities. By doing this it increases the precision of statistics estimated for these groups and gives us a far more accurate insight into the political attitudes and behaviors of these minorities. This enhances the representation and comprehensiveness of the ANES data (Leighley, 2013) [2].

Stratified cluster sampling is used by the ANES as their robust stratified cluster sampling method, a fundamental component of its research design (Lin, 2020) [7]. A simple random sample of Americans would result in a sample of people scattered all over the country, and visiting each one for a face-to-face interview would be prohibitively expensive. Area sampling involves first drawing a sample of regions of the country, and then sampling successively smaller geographic areas within those regions, before sampling specific households within the sampled small areas. The benefit of this procedure is that it is far less expensive to conduct interviews when members of the sample are clustered together in a limited number of areas (DeBell, 2010) [5].

The ANES time series has both panel components and cross-sectional components which provide a comprehensive view of electoral behaviour and political attitudes in the United States and is what makes the ANES data so highly regarded. Interesting statistical explorations have been also carried out like (G. R. Murray, 2010) data mining exploration [8]. The data set has also been used when external factors, such as wars and pandemics affect elections. (L. Baccini,2021) used the data set to identify the effect the COVID-19 pandemic had on the 2020 election [9]. Finally, some key studies done using ANES data that really highlight the possible practical uses of the data set include (Campbell et al., 1980)'s [1] work on "The American voter" which is said to be a key landmark in the study of voting behaviour.

2.1.4 Voter Behaviour (2)

ANES data can also be used to observe any changes and trends in voter behaviour. In the complex landscape of U.S. elections, voter behaviour becomes a fascinating study influenced by economic cycles, social conflicts, and international events. All of which affect the decision of voters when it comes to election time [10]. The American Voter (Campbell et al. [1960] 1980) was a study on the contexts affecting voter behaviour and electoral competition, analysing data from the early American National Election Studies [1]. The study found that American voters have incredibly low levels of information and are largely uninterested in campaigns. However, voters have become more informed and aware as time wore on. (Krosnick and Berent ,1993) found evidence for much higher levels of reasoning than initially visualised. As an example, nowadays international events put foreign policy into focus, affecting a voter's attitude towards the presidential candidates [11]. Voter attitudes towards counter-terrorism, national security and the Iraq and Afghanistan wars were crucial in shaping the presidential elections of the early 2000s [12].

Partisan loyalties amongst the American public have increased greatly since the mid-1970s. A study conducted showed the increase of partisan voting in every presidential election since 1972. By the 1996 election, the level of partisan voting had increased by 77 percent from 1972. These results mirrored the analysis of Miller (1991) and Bartels (1992), who documented the resurgence of partisan voting in the US [13]. A possible explanation for this revival is the civil rights upheavals of the early and middle 1960s (Sundquist 1993, chapter 16; Black and Black 1987; Carmines and Stimson 1989) and viewing the political position of southerners. Party leaders took strong stances on racial issues, with black voters moving towards the left wing and white movers moving towards the right-wing [14].

Move forward to today and partisanship is stronger than ever [15]. The ideological divide between the Democrats and the Republicans is greater than it ever was in the current century (Ansolabehere et al., 2001, Theriault, 2008, Bafumi and Herron, 2010, Mann and Ornstein, 2013, Kraushaar, 2014). The rise of negative partisanship, where voters form political opinions in opposition to political parties they dislike, has led to a big increase in Party loyalty when it comes to voting at all levels. The 81 percent rate of consistent loyalty in the 2012 election was an all-time record [16]. Affect toward political parties as groups has long been recognised as a crucial component of partisanship (Campbell et al., 1960, Green et al., 2002). Democrats and Republicans increasingly dislike opposing party's candidates and also have unfavourable preconceptions about one another's parties [17]. It seems that elite and mass behaviour are mutually reinforcing, which is contributing to the increase of negative partisanship amongst American voters. Republican and Democratic voters are voting along party lines and developing more negative opinions of the other party because of confrontational politics in Washington and many other state capitals. This dynamic interplay between elite and mass behaviour largely responsible for the increase of negative partisanship and the erosion of political trust.

2.1.5 Political Trust (3)

Exploring citizens' sentiments toward politics and government, ANES data emerges as a valuable tool for understanding the trust dynamics. One study by Webster S.W. in 2018 found that anger and politics are strongly connected, anger has a significant influence on how citizens perceive their government parties (Webster, S.W. 2018) [18]. It is logical to believe that citizens tend to trust their government more when the leader aligns with their political ideology, rather than being from a different political background. However, Morisi, Jost, and Singh's research revealed that when the president aligns with one's beliefs, trust is not always uniform among American citizens. Conservative individuals have greater trust in the government when the president shares their political views than liberals do. Also, liberals are more inclined to trust a government led by the opposition (Morisi D., Jost, J., & Singh, V. 2019) [19].

Hollibaugh Jr, G.E. (2016) research show that trust in government goes down when a candidate gathers donations for their political campaign but increases when the candidate is a career-focused individual or an elected member of Congress with expertise in relevant subjects. The research indicates that trust in government falls when nominations are seen as rewarding individuals who promote the financial interests of the administration or the president's party [20].

Cary Wu, Rima Wilkes, and David C. Wilson (2022) argue that trust in the government has hit an all-time low. In 2015, Wilkes analysis of ANES data indicated that Black and White Americans are equally likely to link short-term government performance to trust in government. (Wilkes, Rima 2015). Controversial, Wilkes et al revised this study observing that the decline, spanning since the 1960s, has impacted various demographic groups differently. Notably, the trust gap between racial and ethnic minorities and White individuals has displayed fluctuations both in terms of magnitude and direction. In certain periods, ethnic minorities have exhibited higher levels of trust in contrast to Whites, challenging research assumptions about the link between race and political trust (Cary Wu, Rima Wilkes, David C. Wilson 2022) [21].

While political trust serves as a critical gauge of citizens' sentiments towards the government, it profoundly influences their engagement in various forms of political participation. Understanding the complexities of political trust provides a nuanced backdrop to explore the multifaceted nature of citizens' involvement in democratic processes.

2.1.6 Political Participation (Self)

Political participation in the United States is a multifaceted and vital aspect of its democratic system. Citizens engage in various forms of political involvement, such as voting in elections, participating in political campaigns, attending public rallies and protests, writing to their representatives, and more. However, understanding the complexities of political participation requires a comprehensive approach, as it is influenced by a multitude of factors.

For instance, in Brady et al. (1995) "Beyond Ses: A Resource Model of Political Participation", they emphasised that the motivations of politics alone weren't enough to explain political participation [22]. Instead, using a two-stage least squares estimation, they found that the resources of time, money and skills are also powerful predictors of political participation in the US. Furthermore, they found that political activity can stem from various aspects of individuals' lives including their home, school, jobs and family, and involvement in nonpolitical organisations and churches. They quote from Tocqueville that in this way, this engagement with civil society institutions operates as a "school of democracy."

Blais and Rubenson (2012) contribute to this understanding by highlighting that voter participation has declined steadily in recent years, not only in the United States but across democracies worldwide [23]. Examining political participation across different ethnicities, Ramakrishnan et. Al (2001) found that African Americans were the sole group for which participation increased in a linear manner across generations [24]. In contrast, for White Americans, they found the highest levels of participation to be at the second generation. They suggest that this may be due to "higher levels of cynicism among those in the third generation or higher."

In more recent research, Argyle and Pope (2022) point out a compelling connection between extreme attitudes and political participation [25]. Their study delves into the dynamics of polarisation and identifies specific forms of participation that predict later levels of polarisation. Based on the available data, they conclude that persuasion has the potential to increase the strength of an individual's ideological beliefs although this potential effect is relatively modest.

Understanding political participation requires a comprehensive approach, considering motivations, resources, and the impact of extreme attitudes on later levels of polarisation.

2.1.7 Media Impact on Voting Decisions (4)

For most voters the media is their primary source of information. As a result, it has the power to influence voters' views and opinions. Voters tend to engage with media outlets that align with their beliefs, impacting their understanding of the political scene (Yaser et al. 2011) [26]. Voter behaviour can be influenced by political trust, which is also partially shaped by the media. Unbiased, neutral reporting builds trust while biased reporting can lead to the erosion of it (Moy and Hussain 2011) [27].

In an ideal world, the media would produce unbiased, objective coverage of political issues and well-informed citizens make voting decisions on which candidate will serve their best interests. Unfortunately, this is not the case (Davies, J.J. 2009) [28]. However, the media can never depict the entire truth and is only capable of presenting a partial picture. Competition forces the media to write memorable stories. The best way to accomplish this is to manipulate the facts (Alsem et al. 2008) [29]. If media outlets are seen to favour one party over another, either with a positive one or through providing extensive coverage, this can influence a voter to shift their vote (Cohen and Tsfati 2009) [30]. Research also shows that over exposure or major emphasis on the main campaign issues of one party's campaign has high correlation with how the public vote (McCombs and Shaw 1972) [31]. This is also highlighted in Gavin (1997) where it was found that more prominent issues reported by the media play a key role in the decisions of voters [32]. People's ability to distinguish between fact and fiction has become extremely difficult in today's world as they are being exposed to numerous political rumours and are misinformed by the media (Weeks and Garrett 2014) [33]. This is backed up by Garrett (2011) where it is shown that the more emails received from media outlets containing rumours the more likely someone is to believe in them and consequently share these emails to family and friends [34].

Social media apps such as Twitter and Facebook also play a significant role in influencing voter decisions. If a voter follows a political candidate on any of these apps and is actively engaging with their political discussions, they are more likely to favour that candidate and their party (Biswas et al. 2014) [35]. This is particularly prevalent among younger voters or first-time voters. Young people pay less attention to the traditional sharing of information and are more susceptible to what they see on social media (Intyaswati et al. 2021) [36]. Research shows that during the 2016 elections, 62 percent of US adults got fake news on social media and that the most popular of these fake news stories were widely shared across numerous social media platforms. Most, if not all, of these fake stories favoured Trump, which could explain his victory in the 2016 election (Allcott and Gentzkow 2017) [37]. The Cambridge Analytica scandal was one of the most heavily publicised data breaches of 2018. Data that was illegally collected from unsuspecting Facebook users was used to create tailored advertisements that allegedly aimed to influence voting preferences in the 2016 election (Hinds et al. 2020) [38]. In a study conducted in 2019 on the influences of fake news on Twitter during the 2016 election, 25 percent of stories related to the elections were biased and that Trump supporters are the predominant sharers of these stories.

A multimethodological approach, as advocated by Aday (2010), recognises the need for diverse research methods to comprehensively model the media's effect on voting decisions [39]. This is research that includes the use of more than one method of data collection or research in a study. Biswas et al. (2014) uses both a one-way Anova test and a two-way Anova test to determine if there is a significant relationship between their dependent variables with respect to their independent variables [35]. A binary logit regression model was used by Gunther et al. (2018) to measure the impact of fake news on voting decisions [40]. From their model, it is evident that people who believe in fake news are 3.3 times more likely to sway their vote in favour of the fake news. The media should be considered when trying to predict voting behaviour.

2.2 Focused Literature Review

In this focused section of the literature review, we delve into specific studies that offer detailed insights into the factors influencing voting patterns in past US Presidential Elections. These studies provide nuanced analyses, employing various methodologies to explore the roles of demographics, socioeconomic factors, and political opinions in shaping voter behaviour.

In an analysis of the 2016 US election, Sheng's (2022) paper explores the demographic factors influencing voting behaviour [41]. Sheng utilises logistic regression models, incorporating predictor variables such as race, gender, education, partisanship, age, and income. The study reveals noteworthy findings, including the significant impact of a shift in party affiliation on the likelihood of voting for Trump and the role of income and education in shaping voting preferences. Sheng's models exhibit high sensitivity and specificity, accurately classifying a substantial percentage of cases. However, limitations are acknowledged, such as the omission of media influence and intergenerational transmission in voting behaviour, suggesting avenues for future research.

In a separate study, Zingher's (2020) paper investigates the influence of social class on white vote choice [42]. Zingher develops a class measure based on income, education, occupation, and wealth, finding that lower class standing is a significant predictor of support for Trump. The paper employs a Generalised Structural Equation Model (GSEM) to explore the relationships between class, racial resentment, authoritarianism, and Trump vote choice. Notably, Zingher discovers that while Trump secured a majority vote in the lower class, a substantial portion of his support came from whites in the top half of the class distribution.

Furthermore, Boxell's (2020) paper explores the role of demographic shifts in elucidating recent trends in polarisation [43]. Boxell underscores the increasing educational attainment and ageing of the US population in recent decades, both strongly linked to political involvement. The study delves into the estimation of the relative propensities for polarisation within seven distinct demographic categories, including income, gender, race, religion, education, employment and age. Boxell's findings highlight that the rising levels of education, the growing proportion of elderly individuals, and shifts in religiosity emerge as primary catalysts for demographic-induced alterations in polarisation. Specifically, Boxell suggests that changes in the educational composition of the population can account for 18% of the observed polarization index change, while shifts in religion and age composition contribute an additional 9% and 6%, respectively.

Examining disparities in American voting based on marital status, Weisberg (1987) investigated the impact of being married, in conjunction with considerations of race and income [44]. Weisberg's findings indicate that the marriage gap is predominantly influenced by a combination of racial and income distinctions. Notably, in 1984, Weisberg observed that married individuals were voting for Republicans at a rate 10-15% higher than their single counterparts. The author acknowledges that the marital gap stands apart from more common demographic voting differences, asserting that marital status disparities stem from decisions and events spanning the teenage years to middle and old age. Consequently, Weisberg raises the issue of whether the marriage gap merely mirrors other distinctions in voting behaviour.

In a paper exploring immigrant threat and voter choice in the 2016 US election, Garand et. al (2020) examine the effects of immigrant threat perception with American voter identity, following Trump's emphasis on the "importance of American identity (e.g., "America First," "Make America Great Again")" [45]. In a logistic regression model, examining Americans' perceptions of immigration cultural, crime, economic, and general immigrant threat, it was found that individuals who have a strong sense of American identity are strongly and significantly more likely to have a general sense of immigrant threat (b = 0.099, z = 6.14). In a second logistic regression model estimating the effects of immigration threat and American identity on vote choice in the 2016 presidential election with Trump encoded as 1 and Hillary Clinton as 0, American Identity, Immigrant threat scale and the interaction term between these two were all found to be statistically significant. From this, we interpret that not only does American Identity have a strong effect on perceptions of immigrant threat, but it also has a strong effect on vote choice. The statistical significance of the interaction term indicates that the effect of immigration threat on presidential vote choice in the 2016 presidential election was enhanced for individuals with a strong American identity.

In a paper examining populism in the 2016 US election, populist language scores were assessed for the leading seven candidates (Oliver and Rahn, 2016) [46]. While examining language simplicity, Trump had the lowest average words per sentence with also having the lowest variety in his language. This easily understood language in turn made Trump more "of the people" contrary to some strictly economically populist candidates such as Bernie Sanders who had a higher average words per sentence at 21.38 and language variance of 0.29. The paper concludes that the emergence of Donald Trump was ultimately rooted in American party politics. The paper is closed out by noting that modern American populism has a "conservative tinge" and is felt most in the Republican Party.

In a paper exploring abortion opinions and voter choice in the 1992 US Presidential Election, Abromowitz (1995) found abortion to have the strongest influence on vote choice than all other policies included in the study, including social welfare, affirmative action, defence spending, the Gulf War and the death penalty [47]. The study, compiling data similarly from the ANES, found that voters did not stick strictly to their affiliated party's policy. Just 11% of Republicans held the party's "pro-life" view of never permitting abortion while 31% of Democrats had a view conflicting with the party's policy at the time. Abromowitz concludes that abortion had a much greater impact on the Republican party than the Democratic party as those who were "aware and concerned" about this issue were disproportionately white, affluent and well-educated. He claims that the Democratic party suffered less from pro-life Democrats as most of these Democrats either "didn't know the candidates' positions on abortion or didn't care about the issue"

2.3 Summary

Expanding on this existing literature, my analysis will delve into demographic, socioeconomic, and political opinion variables. Initially, I intend to conduct an exploratory analysis to provide a comprehensive understanding of each variable's impact. Subsequently, I will employ both multiple logistic regression and random forest modelling techniques to compare distinct categories against voter choice. This multifaceted approach aims to offer a thorough examination of the factors influencing electoral decisions.

Chapter 3

Exploratory Data Analysis

Overview

In this section, we explore the key characteristics of the ANES 2020 Time Series dataset, with a focus on demographic, socioeconomic and political opinion variables. The dataset, sourced from the ANES website, is titled '2020 Time Series Study' and comprises 8280 entries with 1771 total columns [48]. Each column is identified by a reference code (e.g., "V212345"), cross-referenced in the ANES 2020 User Guide and Codebook [49]. For the purposes of this paper, we are only examining those who voted for Trump or Biden, leaving us with 5729 entries. Demographic and socioeconomic variables of interest in this analysis include sex, age, ethnicity, social class, state of registration, rural/urban living, Liberal/Conservative identity, religion, marital status, education and income. Political opinion variables of interest in this analysis are stance on abortion, the death penalty, defence spending and immigration policy.

Note: Responses such as "Refused", "Don't know", and "Interview Breakoff" will be excluded from graphs unless deemed important. In graphs where voting patterns are illustrated, red represents those who voted for Trump while blue represents those who voted for Biden. In certain cases, graph labels were abbreviated, with the complete description available within the accompanying R Code.

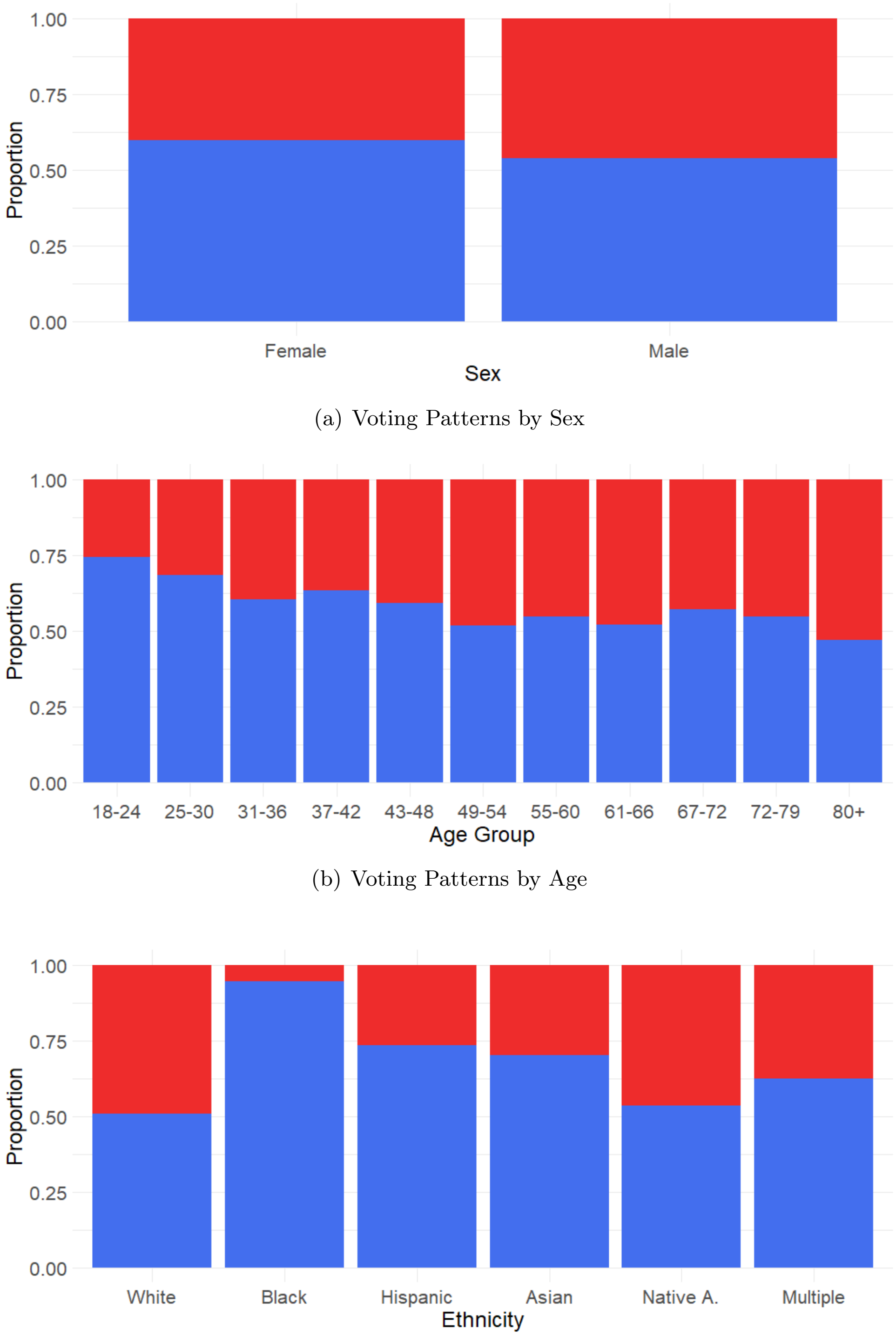
3.1 Demographic Variables

This section delves into key demographic variables — sex, age, and ethnicity. Figure 3.1 provides a visual representation of voting patterns based on these factors, using ANES labelling conventions for consistency.

Sex: Just under 55% of respondents are female, while just under 45% are male. The graph suggests no clear difference in preferences based on sex; however, females exhibit a slightly higher preference for Biden compared to males.

Age: The average respondent age is 53 years, with a standard deviation of 17 years. Examining voting patterns by age reveals intriguing insights. Younger respondents, notably the 18–24-year-old age group demonstrate a considerable preference for Biden, while those aged 49 and above exhibit a more balanced preference.

Ethnicity: White Americans contribute significantly, constituting 76% of the dataset, aligning closely with the estimated population parameter of 75.5% as of July 1, 2022, according to the United States Census Bureau [50]. Figure 3.1 illustrates that White Americans and Native Americans show nearly equal support for Biden and Trump. Conversely, the Black community overwhelmingly favours Biden, with proportions of 94% to 6%, respectively. Those of other ethnicities also demonstrate a clear preference for Biden.



(a) Voting Patterns by Sex

(b) Voting Patterns by Age

(c) Voting Patterns by Ethnicity

Figure 3.1: Voting Patterns by Demographic Variables.

3.2 Socioeconomic Variables

Investigating the socioeconomic landscape in this analysis involves delving into key variables such as social class, education, income, religion, marital status, liberal/conservative ideology and living area, as illustrated in Figure 3.2 - 3.4 showcasing voting patterns.

Social Class: Previous studies have underscored the significance of social class in predicting voter choices. Analysis of voting patterns reveals a balance between Biden and Trump preferences within the working and middle class, while the upper and lower class lean towards Biden.

Education: Turning to education, those with Bachelor's or Graduate degrees exhibit a preference for Biden, while those with a 'High school credential' show a slight preference for Trump (54%).

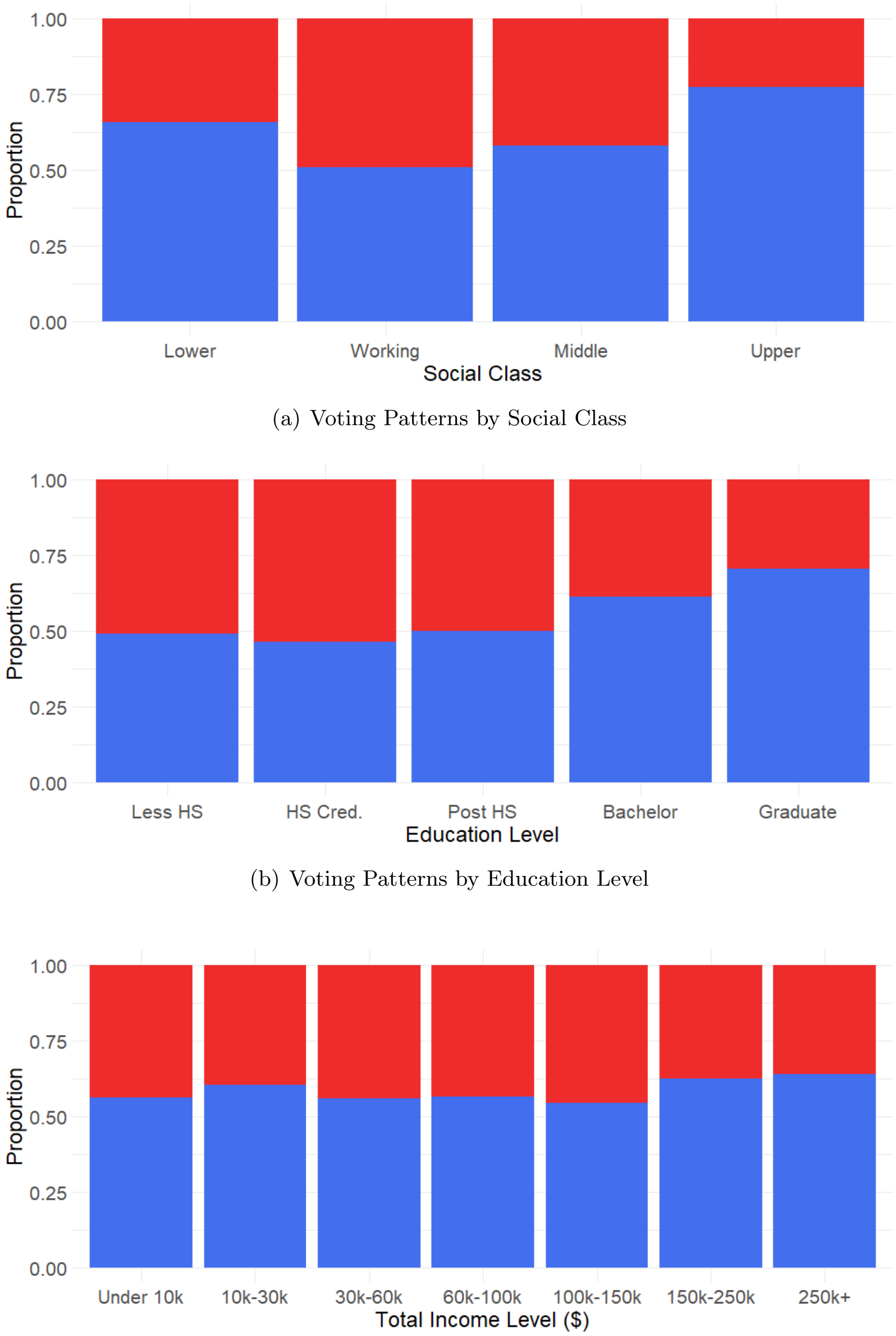
Income: Remarkably, this exploratory analysis indicates a lack of noticeable difference in preferences between Biden and Trump, except for individuals earning between \$10,000 and \$30,000, and those earning \$150,000 or more, among whom a slight preference for Biden is apparent.

Religion: Religious affiliations have been consolidated into a category labelled 'Non-Religious,' including individuals identifying as Atheist, Agnostic, Something else, or Nothing in particular. Among specific religious groups, Latter Day Saints (LDS) and Protestants lean towards Trump, whereas Roman Catholics and Orthodox Christians exhibit an equal preference for both candidates. Respondents from diverse religious backgrounds demonstrate a notable preference for Biden, mirroring patterns observed in ethnic preferences.

Marital status: Marital status reveals that the majority of respondents are married (56%). Among the married, widowed, and divorced groups, preferences for Biden and Trump are approximately equal. In contrast, those who are separated or have never been married show a stronger preference for Biden.

Liberal/Conservative Ideology: Examining voting patterns by liberal/conservative ideology uncovers intriguing trends. Conservatives only slightly favour Trump, while Liberals overwhelmingly favour Biden. A significant proportion (69%) finds the question inapplicable, displaying nearly no preference. Notably, both Liberals and Conservatives received a roughly equal number of votes.

Living area: Finally, looking at voting preferences by living area, 'Rural area' leans towards Trump, while 'Suburb' and 'City' show a strong preference for Biden.

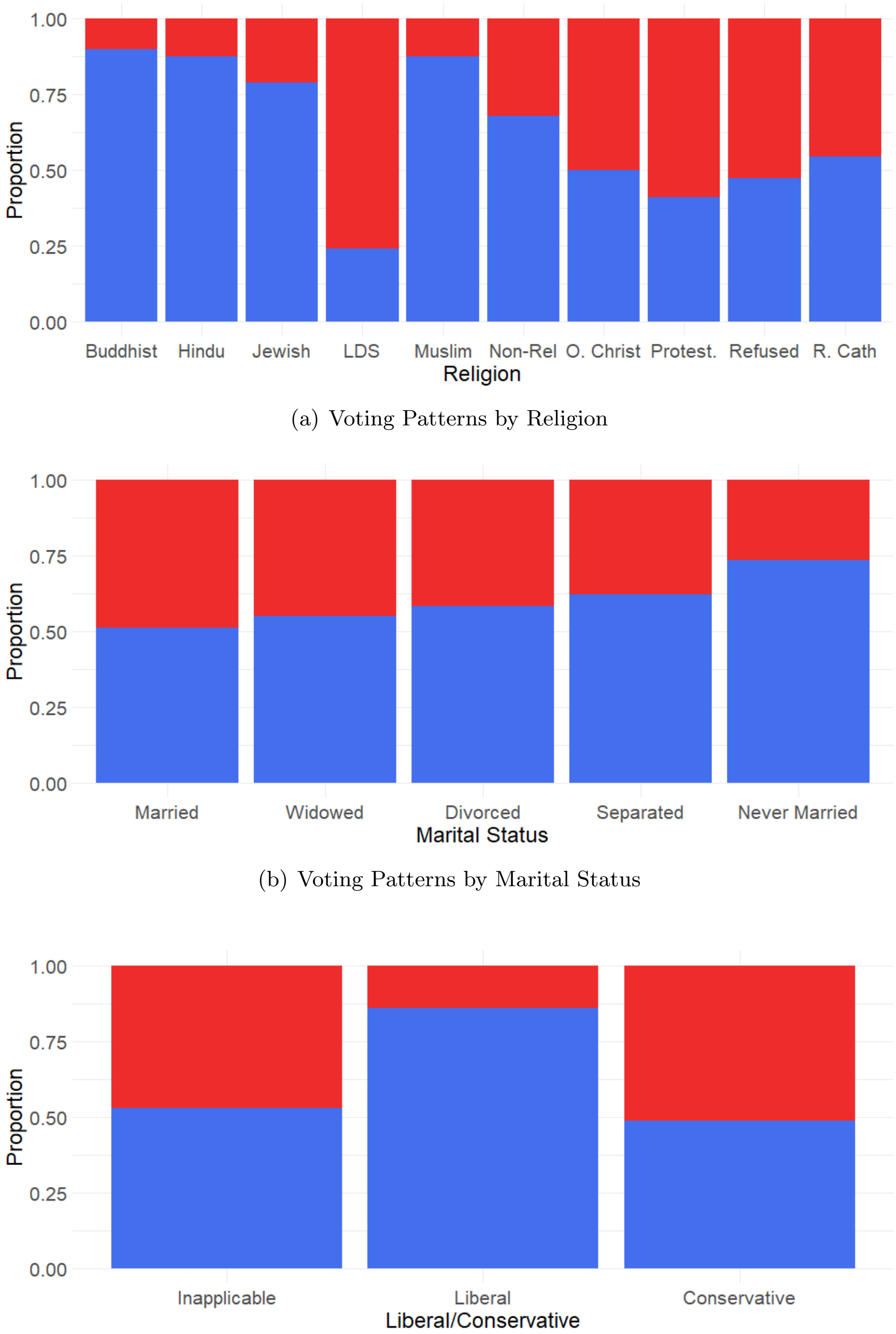


(a) Voting Patterns by Social Class

(b) Voting Patterns by Education Level

(c) Voting Patterns by Income

Figure 3.2: Voting Patterns by Socioeconomic Variables.

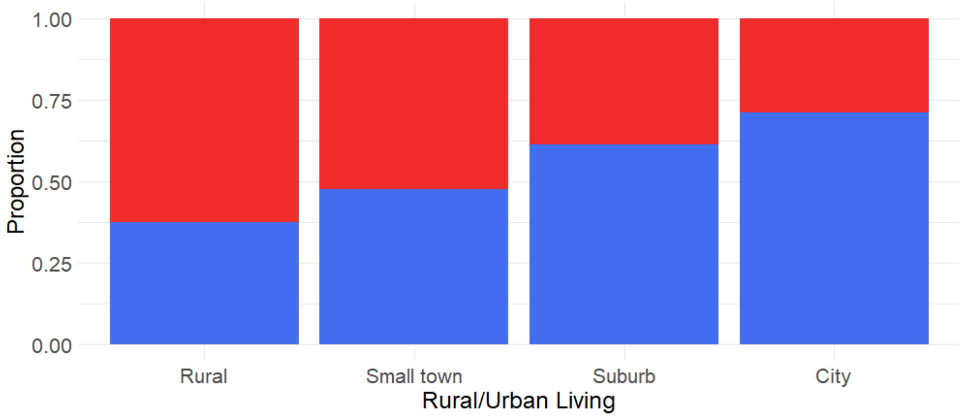


(a) Voting Patterns by Religion

(b) Voting Patterns by Marital Status

(c) Voting Patterns by Liberal/Conservative

Figure 3.3: Voting Patterns by Socioeconomic Variables.



(a) Voting Patterns by Rural/Urban Living

Figure 3.4: Voting Patterns by Socioeconomic Variables.

3.3 Political Opinion Variables

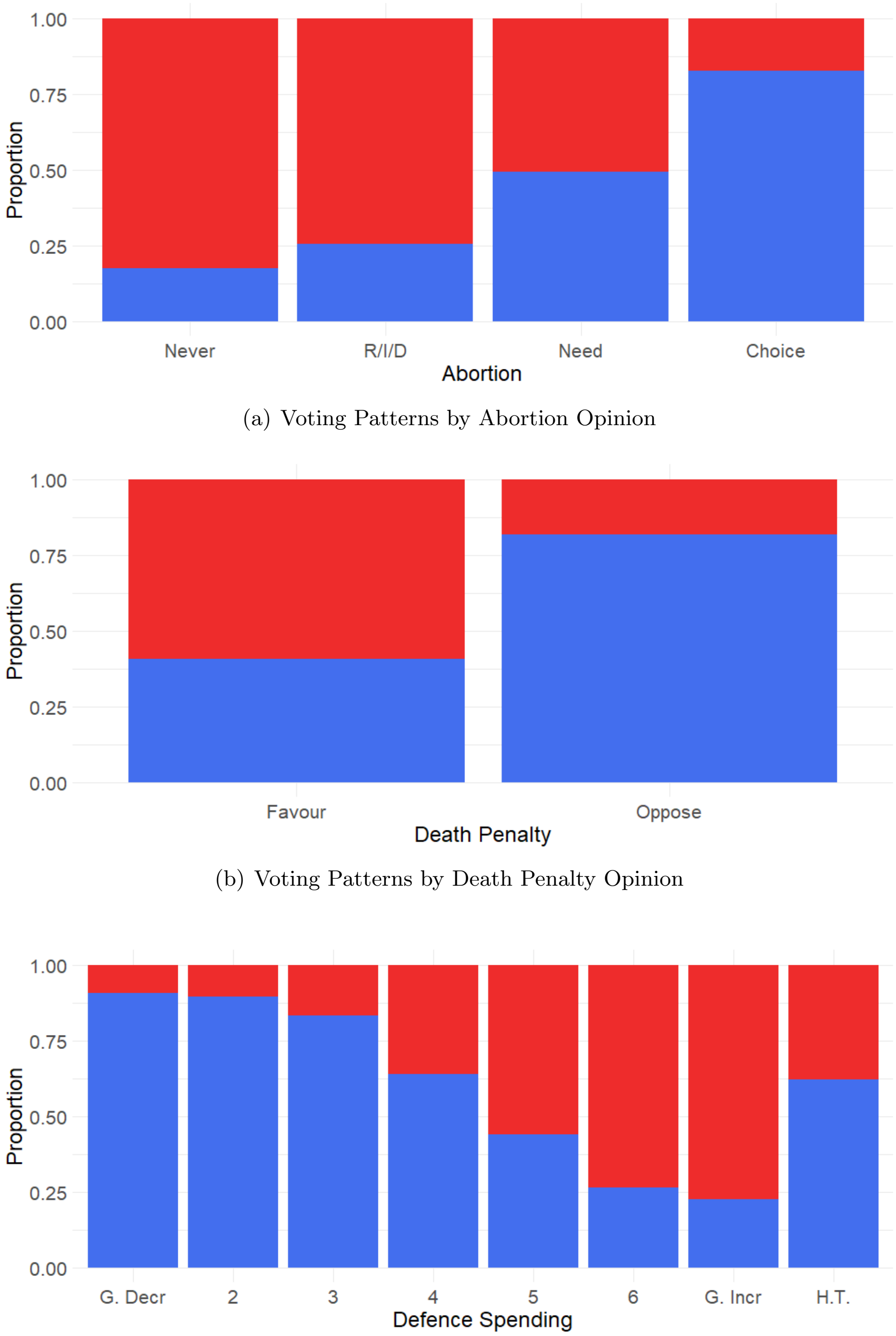
In this section, we delve into respondents' perspectives on key political issues, examining their stances on controversial questions and exploring how these opinions correlate with their voting choices, as illustrated in Figures 3.5 and 3.6 showcasing voting patterns.

Abortion: The contentious topic of abortion reveals a diverse range of opinions among respondents. Almost half (49%) advocate for a woman's right to obtain an abortion as a matter of personal choice, while just over 10% firmly believe that abortion should never be legally permitted. Charting the voting patterns concerning these views unveils a compelling trend—a pronounced preference for Trump among those opposing any legal permission for abortion, gradually shifting towards a stronger preference for Biden as restrictions on abortion decrease.

Death Penalty: When queried about their stance on the death penalty for individuals convicted of murder, approximately 60% of respondents expressed support, with just over 39% opposing. Examining the voting patterns tied to these perspectives reveals a distinct divide—those in favour of the death penalty predominantly voted for Trump, while a larger majority of those opposed to it cast their votes for Biden.

Defence Spending: Respondents were asked to position themselves on a 7-point scale ranging from greatly decreasing to greatly increasing defence spending. A notable portion (10%) admitted to not having given much thought to this issue. Meanwhile, just over 10% positioned themselves at the extreme upper end of the scale, over 21% directly in the middle, and over 8% at the bottom. Analysing voting patterns exposes a parallel trend to the abortion issue, with a strong Biden preference among those at the bottom of the scale, gradually shifting towards a more pro-Trump inclination as the scale progresses. Intriguingly, respondents who haven't contemplated this matter exhibit only a slight Biden preference.

Immigration: Immigration has been a highly debated and heated topic in the US in recent years. Respondents were given four different options on what government policy should be toward unauthorised immigrants now living in the US. Over half of respondents (56%) had the view that unauthorised immigrants should be allowed to remain in the US and eventually qualify for citizenship but only if they, met certain requirements. Over 12% of respondents would like the government to make all unauthorised immigrants felons and send them back to their home country, while a further 17% believe the government should allow all unauthorised immigrants to remain in the US & eventually qualify for citizenship without penalties. Examining voting patterns unearths a similar trend seen in the abortion and defence spending issue with a heavy Trump vote inclination at the harshest restrictions on immigration, gradually shifting to a stronger Biden preference as immigration restrictions decreased.

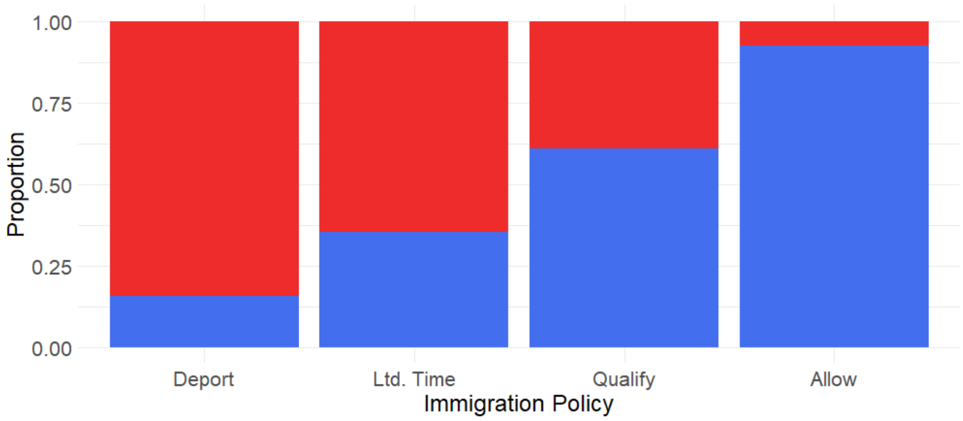


(a) Voting Patterns by Abortion Opinion

(b) Voting Patterns by Death Penalty Opinion

(c) Voting Patterns by Defence Spending Opinion

Figure 3.5: Voting Patterns by Political Opinion Variables.



(a) Voting Patterns by Immigration Policy

Figure 3.6: Voting Patterns by Political Opinion Variables.

3.4 Summary

This exploratory analysis scrutinises the ANES 2020 Time Series dataset, uncovering insightful patterns in voter behaviour. Among the key findings are distinct voting preferences within demographics such as age, ethnicity, and social class. Notably, while White and Native American respondents exhibit a balanced split between Biden and Trump support, the Black community overwhelmingly favoured Biden (94% to 6%). Socioeconomic factors, including education and living area, underscore intriguing trends, with higher educational attainment correlating with a pro-Biden inclination. Furthermore, the exploration of political opinion variables uncovers compelling connections; for instance, respondents opposing legal permission for abortion display a pronounced preference for Trump, gradually shifting toward a stronger preference for Biden as restrictions on abortion decrease. These specific findings provide a nuanced understanding of the intricate dynamics shaping the electoral landscape during the 2020 US election.

Chapter 4

Mathematical Modelling

In this paper, a Multiple Logistic Regression and Random Forest model are used to predict a binary outcome; a vote for Trump or Biden, based on a multitude of predictor variables. In this section, we will explore these methods and analyse how they work. We will also examine how these models can then be evaluated at the end of the section using evaluation techniques such as k-Fold Cross-Validation and different performance metrics.

4.1 Multiple Logistic Regression

[51] Multiple Logistic Regression builds on Logistic Regression which ultimately builds on Linear Regression, therefore we will start there. Simple Linear Regression aims to predict a response variable Y from one predictor variable X, assuming a linear relationship between the two. This can be written mathematically as:

$$Y \approx \beta_0 + \beta_1 X \tag{4.1}$$

where β_0 and β_1 are known as the coefficients/parameters. A simple linear regression model aims to fit a line to the data well. β_0 is known as the intercept and β_1 the slope. These parameters are estimated most commonly by minimising the **least squares** criterion. If we let $\hat{y_i} = \hat{\beta_0} + \hat{\beta_1} x_i$ be the predicted value for our response variable Y based on the i$^{\text{th}}$ value of X, then the i$^{\text{th}}$ residual, $e_i = y_i - \hat{y_i}$, defines the difference between the i$^{\text{th}}$ observed value and the i$^{\text{th}}$ predicted value. We now define the Residual Sum of Squares (**RSS**) as:

$$RSS = e_1^2 + e_2^2 + ... + e_n^2 \tag{4.2}$$

Multiple Linear Regression simply builds on this by adding more predictor variables. The equation can be written as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_j X_j + \epsilon \tag{4.3}$$

where X_j represents the j$^{\text{th}}$ predictor variable, β_j represents the coefficient associated with the j$^{\text{th}}$ predictor variable and ϵ represents the error in the estimate. The coefficients are again found by minimising the least squares criterion.

[51] Now that we have looked at Simple and Multiple Linear Regression, we can look at Logistic Regression. Simple Logistic Regression is used to predict a binary outcome from one predictor variable. The model predicts a certain probability that Y belongs to a particular category. Naturally, as we are predicting a probability, we must ascertain our prediction to lie between 0 and 1. To do this, we use the **logistic function:**

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \tag{4.4}$$

Unlike Linear Regression, **maximum likelihood** (ML) is used to find the coefficients. Least squares can be used here but ML is generally preferred due to its superior statistical properties. The intuition behind ML is to find a β_0 and β_1 so that when we get an estimate for $p(x)$, it is close to 1 for all successes and conversely close to 0 for all failures. The likelihood function for a Bernoulli random variable can be formalised as follows:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) \tag{4.5}$$

The estimates for β_0 and β_1 are chosen to maximise the likelihood function.

From manipulating (4.4) we obtain the following:

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X} \tag{4.6}$$

The quantity on the LHS is known as the **odds**, ranging from 0 to ∞. Values close to 0 indicate very low probability whereas those larger close to ∞ indicate very high probabilities. If we continue and take the natural logarithm of both sides of (4.6) we obtain:

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X \tag{4.7}$$

The LHS of this equation is known as the **log odds** or **logit**. We can see that (4.4) has a logit that is linear in X.

By exponentiating a coefficient β_j, we can obtain an **odds ratio**. Odds ratios are used to interpret the coefficients. The odds ratio compares the odds of an outcome occurring in one group compared to another group.

Building on this, Multiple Logistic Regression predicts a binary response using multiple predictors. We can generalise (4.7) as follows:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_j X_j \tag{4.8}$$

where $X = (X_1, ..., X_j)$ are j predictor variables. Furthermore (4.8) can now be rewritten as:

$$p(x) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_j X_j}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_j X_j}} \tag{4.9}$$

Similar to Simple Logistic Regression, we use the maximum likelihood function again to estimate the coefficients $\beta_0, \beta_1, ..., \beta_j$.

From the predicted values we obtain, a default threshold of 0.5 will decide a success or failure by assigning all values strictly below this value as failures and conversely, any values above or equal to this value will be assigned a success. Methods such as examining **Receiver Operating Characteristic** (ROC) curves can be used to identify a suitable threshold by finding a suitable trade-off between the true positive and false positive rate. A different threshold may also be set earlier on if for example there is a threshold on the number of false positives the researcher can allow[51].

4.2 Random Forest

To first understand how a Random Forest works, one must understand the fundamentals of a Classification and Regression Trees (CART) algorithm. CART can be used for both regression and classification tasks, here it is being used as a classifier.

To begin, we must first understand some terminology of a tree. A tree is usually seen in an inverted form, with the **root** node being at the top. From the root node, we can either go down left or right, typically left means the node is true and conversely right is false. The next node we come onto is known as an **internal** node. Internal nodes are inside of a decision tree in which the preceding node branches out into two or more variables. For this example, we will consider just two outcomes (true or false). A **Parent** node is a node that has at least one node below it, known as a **child** node. The lines connecting all of these nodes are known as **branches**. At the end of a tree, we have a **leaf** or **terminal** node. These nodes have branches pointing to them but nothing after them [52]. Figure 4.1 below displays a simple tree for interpretation.

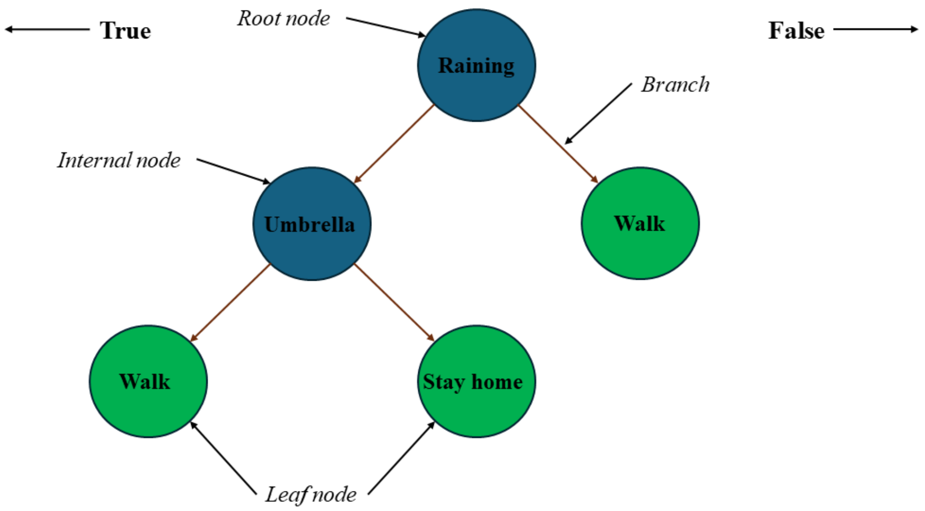


Figure 4.1: A simple decision tree illustrating basic terminology. In this example, the decision tree determines whether one will go for a walk based on rainfall. By utilising true and false conditions at each node, a final decision is reached.

[51] Now that we have examined the basic structure of a decision tree, we must understand a few terms. We must decide which predictor variable (often referred to as a feature) should we choose to start off the tree with. This is known as **feature selection**. For classification trees, we aim to minimise **impurity** and for regression, we aim to minimise **variance**. In this paper, we will just explore classification trees as that is the form of our model. In general, impurity can take a value from 0 to 1. A node with its impurity $= 0$ or 1 is said to be **pure**. Purity measures the distribution of true/false outcomes within each child node from the parent. Ideally, all leaf nodes should be pure for decision-making. For classification trees, the classification error rate can be used for feature selection. It can be defined as:

$$E = \max_k(\hat{p}_{mk}) \tag{4.10}$$

where $\hat{p}_{mk}$ represents the proportion of training observations in the m$^{\text{th}}$ region from the k$^{\text{th}}$ class. This approach is not sufficiently sensitive for tree-growing and other approaches are preferable. In this paper, we will explore the **Gini Index** as that is what is used in R. It can be defined as:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{4.11}$$

where G is defining a measure of total variance across the K classes. In this paper, K = 2, Trump or Biden. From the sum, it can be intuitively understood and proven that G will take a smaller value the closer all of the $\hat{p}_{mk}$'s are to zero or one. For our case, we note that it will reach its maximum value when $\hat{p}_{mk} = 0.5$, where a Trump and Biden outcome are equally likely.

Now that we have all these terms we can explore how a CART algorithm works. We will look at a classification tree for the purposes of this paper.

Binary Splitting: CART starts by selecting the best feature and its optimal split point to divide the dataset into two subsets. Choosing the best split involves minimising the impurity.

Recursive Partitioning: Once the dataset is divided into two subsets, CART repeats the splitting process on each subset. This process continues recursively until a stopping criterion is met. This criterion might involve reaching a maximum tree depth, having a minimum number of samples in each leaf node, or other conditions.

Tree Pruning: After the tree is fully grown, it might be pruned to avoid overfitting. Pruning involves removing nodes that do not provide significant improvements to the model's performance on a validation dataset.

Prediction: Once the tree is constructed, predictions are made by traversing the tree from the root node down to a leaf node. For classification tasks, the majority class in the leaf node is assigned as the predicted class [53].

Now that we have established how a CART algorithm works, we can examine a Random Forest. The Random Forest aims to assign classes to observations using a set of predictors. Random Forest splits the data into training and testing datasets. Usually, an 80/20 or 90/10 split of training to testing data will be taken. The training data is known as the "in-bag" data, while the testing data is known as the "out-of-bag" (OOB) data. Moreover, the OOB error or estimate is a method for measuring the prediction error of the model. Formally, the random forest classifier can be represented as a collection of tree-structured classifiers $\{f(\mathbf{x}, \Theta_k), k = 1, \ldots\}$, where $\mathbf{x}$ is an input vector of P predictor values, and k is an index for a given tree. Each Θ_k is a random vector constructed for the k^{th} tree so it will be independent of previous random vectors $\Theta_1, \ldots, \Theta_{k-1}$, and are all generated from the same distribution. This describes how subsets of predictors are sampled without replacement for each potential split. The algorithm for performing a Random Forest with N observations - is as shown below.

1. Take a random sample of size N with replacement from the data.
2. Take a random sample without replacement of the predictors.
3. Construct the first CART partition of the data.

4. Repeat Step 2 for each subsequent split until the tree is as large as needed. Do not prune.
5. Run the out-of-bag data through the tree. Store the class assigned to each observation along with each observation's predictor values.
6. Repeat Steps 1–5 many times (typically 500).
7. Utilising only the class assigned to each observation when that observation is excluded from tree construction, count the number of times over all trees that the observation is classified in one category and the number of times it is classified in the other category.
8. Assign each case to a category by a majority vote over the set of trees [53].

R has an importance function which ranks the importance of each predictor variable used in the model. For a classification tree, R uses the mean decrease in impurity (MDI) of nodes measured by the Gini Index. The algorithm for finding a feature's importance is shown below.

1. Construct a measure prediction error v for each tree by running the out-of-bag error rate (OOB) through the tree.
2. If there are p predictors, iterate Step 1 p times, each time shuffling the values of the features randomly. This randomisation ensures that, on average, each predictor is uncorrelated with the response variable and the other predictors. Subsequently, for each shuffled predictor j, new metrics for prediction error, denoted as v_j, are computed.
3. For each of the p predictors, calculate the average difference across all trees between the prediction error without shuffling and the prediction error when the j^{th} predictor is shuffled.

It follows that the average increase in forecasting error when a given predictor j is shuffled represents the importance of that predictor at forecasting. The formula is as follows:

$$I_{\text{j}} = \sum_{k=1}^{K}[\frac{1}{K}(v_{\text{j}} - v)],\ j = 1, \dots, p \tag{4.12}$$

where K is the number of trees, v_j is the forecasting error with predictor j shuffled and v is the forecasting error with none of the predictors shuffled. It should be noted that forecasting accuracy may improve slightly when a variable is shuffled due to the randomness introduced. A negative in forecasting importance can be treated as no decline in accuracy and is often simply ignored. The preferred measure for the forecasting errors, v and v_j, is the proportion of cases misclassified [53].

4.3 Model Evaluation Techniques

In this paper, we will look at a multitude of model evaluation techniques. In this section, we will examine some of the main techniques that will be used later in the paper.

4.3.1 Performance Metrics

Due to the binary nature of our response variable, our classification output must come out under one of these four categories [54].

1. True positive (TP)
2. True negative (TN)
3. False positive (FP)
4. False negative (FN)

These categories can help us identify the performance of our classifier. We can now identify many paramaters to evaluate performance. These are:

1. **Sensitivity/Recall:** The sensitivity/recall of a classifier is the ratio between how many observations were correctly classified as positive to how many were actually positive. This can be written as $\frac{TP}{TP+FN}$
2. **Specificity:** The specificity of a classifier is the ratio between how many observations were correctly classified as negative to how many were actually negative. This can be written as $\frac{TN}{TN+FP}$
3. **Precision:** The precision of a classifier is the ratio between how many observations were correctly classified as positives out of all positives. This can be written as $\frac{TP}{TP+FP}$
4. **F1 score:** The F1 score of a classifier is the harmonic mean of precision and recall which is a measure of the model's classification ability. The F1 score is widely considered a better indicator of a classifier's performance rather than the usual accuracy measure. This can be written as

$$2\left(\frac{precision * recall}{precision + recall}\right) \tag{4.13}$$

4.3.2 K- Fold Cross Validation

[52] As a part of our model evaluation techniques, we will look at a method known as **k-Fold Cross-Validation** (k-Fold CV). As the modelling methods in this paper are for classifying observations, we shall look at k-Fold CV for classification. K-Fold CV works by dividing the dataset into k groups (folds), of approximately equal size. We treat the first fold as the validation set, the modelling method is then fit to the remaining $k-1$ folds. The error rate (using the number of misclassified observations) is then computed on the validation set. This process is then repeated k times, using a different fold as the validation set each time. This then results

in k estimates of the error rate, $Err_1, Err_2, ..., Err_k$. The k-Fold CV estimate is hence found by averaging these values, formally writing as:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} Err_{(i)} \tag{4.14}$$

In practice, it is most common to use a value of $k = 5$, or $k = 10$ [52]. Note that when we evaluate our models later using k-Fold CV, R provides additional information with Random Forest, displaying the accuracy values for the number of variables tried at each split.

4.3.3 Akaike Information Criterion

The Akaike Information Criterion (AIC) is a statistical information measure which balances the goodness of fit in the model along with the number of parameters. Akaike proposed the use of the Kullback-Leibler (K-L) divergence equation (a metric used to compare two data distributions), as the fundamental basis for model selection. It was found that the empirical log-likelihood function at its maximum point could estimate K-L information. In estimating the model selection target, the maximised log-likelihood was found to be biased upwards. Under certain conditions, Akaike found that this bias was approximately equal to K, the number of parameters in the model.

In producing the seminal equation, Akaike multiplied $log(\mathcal{L}(\hat{\theta}|y)) - K$ by -2 for “historical reasons”. The equation can be defined as follows:

$$AIC = -2log(\mathcal{L}(\hat{\theta}|y)) + 2K \tag{4.15}$$

where $\hat{\theta}$ are the estimated model parameters and y is the observed data.

The AIC is on an interval scale meaning we examine absolute differences rather than percentages. Lower AIC values indicate better model fit and parsimony, making them preferable for model selection. In this paper, the AIC will naturally be used for model selection on our multiple logistic regression model, finding predictor variables that are not contributing significantly and thus removing them, lowering the AIC [55].

Chapter 5

Modelling Methods

This paper uses Multiple Logistic Regression and Random Forest models to explore the predictive power of demographic, socioeconomic, and political opinion variables in determining a vote for either Trump or Biden. The models incorporate all variables explored in the EDA. Note that the levels labelled “Not Specified”, entail levels such as “Interview Breakoff” or “Technical error”. In some cases, the “Refused” category is also put into this level unless there is a substantial amount in that category.

5.1 Multiple Logistic Regression Model

In this model, Biden is encoded as the dummy variable, focusing the analysis on voter preferences for Trump over Biden. Initially, all variables explored in the EDA were included in the model. When looking at model reduction, income and education were removed from the model for having weak differences to the AIC when removed (< 3) [55]. The importance was found by the change in AIC when the variable was removed from the model against the full model AIC.

Figure 5.1 plots the variables by decreasing in importance for the final model. From the figure, we see that immigration policy has the biggest effect on AIC followed by abortion, defence spending, liberal/ conservative identity and ethnicity. The importance takes a sharp drop going onto the death penalty and again after religion. The importance of the last five variables is quite low in comparison to the remainder. Following this, Table 5.1 presents the summary output from R of all variables used in this new model, listed in order of importance.

Examining Table 5.1, we can interpret the coefficients using the odds ratios mentioned in Section 4.1 by exponentiating the coefficients. In particular, we can say that:

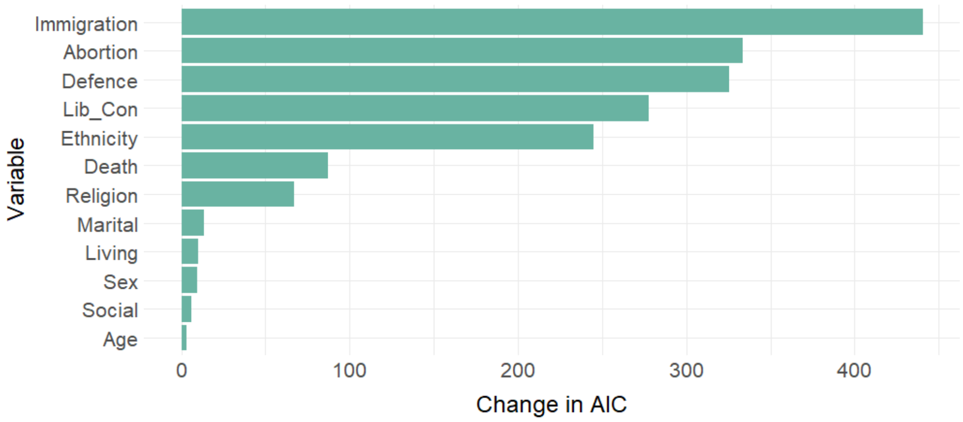


Figure 5.1: Logistic Importance of the final model ranked from top descending via the change in AIC.

- Those who would like the US govt. to make all unauthorised immigrants felons and send them back to their home country are over 30 (30.57) times more likely to vote for Trump than those who would like to allow all unauthorised immigrants to remain in the US and eventually qualify for citizenship without penalties. This finding affirms our EDA earlier on immigration policy in which we saw Trump voters tended to be a lot more restrictive on this matter compared to Biden voters.

- Examining abortion, those who believe abortion should never be permitted were approximately four and a half (4.53) times more likely to vote for Trump than those who believe it should be a matter of personal choice. Noticing the two significant levels below these, we can see that the coefficients remain positive, further affirming that Trump voters tend to favour stricter restrictions on abortion than Biden voters.

- For defence spending, those who would like the US govt. to greatly increase defence spending were over 10 times (10.28) more likely to vote for Trump than those who would like to greatly decrease government spending on defence.

- For those who would label themselves a Conservative, they were almost 5 times (4.85) more likely to vote for Trump than those who would identify as Liberal. Interestingly, those who believed this question was inapplicable (69%), were over 8 times (8.24) more likely to vote for Trump over liberals. This may suggest that Liberals would have a strong association with being a Biden supporter whereas Conservatives may not be strictly associating themselves with being a Trump supporter.

- Looking at ethnicity, an immediately noticeable category is Black Americans, who were approximately 20 times less likely (0.05) to vote for Trump than White Americans. We saw this early preference in the EDA where 94% of Black Americans voted for Biden in our dataset.

Variable	**Estimate**	**Variable**	**Estimate**
(Intercept)***	−6.09	Death.Favour***	0.83
Immigration.Deport***	3.42	Death.fNS	0.28
Immigration.LimitedTime***	2.39	Religion.Buddhist*	−1.20
Immigration.Qualify***	1.47	Religion.Hindu*	−1.65
Immigration.NS**	1.79	Religion.Jewish	−0.37
Abortion.Never***	1.51	Religion.LDS***	1.13
Abortion.Cases***	0.90	Religion.Muslim.	−1.70
Abortion.Need***	0.43	Religion.O. Christian*	0.81
Abortion.Other Opinion***	−0.91	Religion.Protestant***	0.68
Abortion.NS	0.80	Religion.R. Catholic	0.16
Defense.gdplus1	−0.33	Religion.NS	−0.01
Defense.gdplus2	0.22	Marital.Married***	0.47
Defense.gdplus3***	0.99	Marital.Divorced	0.10
Defense.gdplus4***	1.51	Marital.Separated	−0.05
Defense.gdplus5***	2.20	Marital.Widowed	0.27
Defense.Greatly Increase***	2.33	Marital.NS	0.56
Defense.Haven't Thought***	1.18	Living.Rural***	0.46
Defense.NS	1.49	Living.Small Town**	0.34
Lib_Con.Conservative***	1.58	Living.Suburb	0.10
Lib_Con.Inapplicable***	2.11	Living.NS	0.99
Lib_Con.NS***	1.20	Sex.Male***	0.29
Ethnicity.Asian.	−0.38	Sex.NS	0.72
Ethnicity.Black***	−2.96	Social.Lower	−0.06
Ethnicity.Hispanic***	−0.84	Social.Middle	0.28
Ethnicity.Multiple**	−0.60	Social.Working*	0.50
Ethnicity.Native American	0.34	Social.NS	−0.29
Ethnicity.NS	−0.67	Age *	−0.01

Table 5.1: Summary of Logistic Regression Model. AIC = 4226.9. p-values: . $< .1$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Other notable findings include but are not limited to; married people were 60% (1.60) more likely to vote for Trump than those who were never married, men were nearly 34% (1.34) more likely to vote for Trump than women and those who favour the death penalty for murder were over twice (2.29) as likely to vote for Trump than those who opposed it.

5.2 Random Forest Model

In this section, we examine a Random Forest model used to explore the predictive power of demographic, socioeconomic, and political opinion variables in determining a vote for either

Biden or Trump. The model incorporates all variables explored in the EDA. In this model, due to Biden having more votes than Trump (3267 to 2462), the classes are weighted 1.75 to 2.33 respectively. The model is run with 500 trees, trying 2 variables at each split. This value was chosen from running 10-Fold Cross-Validation on the model, yielding an accuracy (81.6%) similar to the best accuracy obtained from using 34 variables (82.9%). Using 2 variables instead will provide a simpler model which will be easier to interpret and also less computationally expensive. The number of trees used was chosen from examining Figure 5.2.

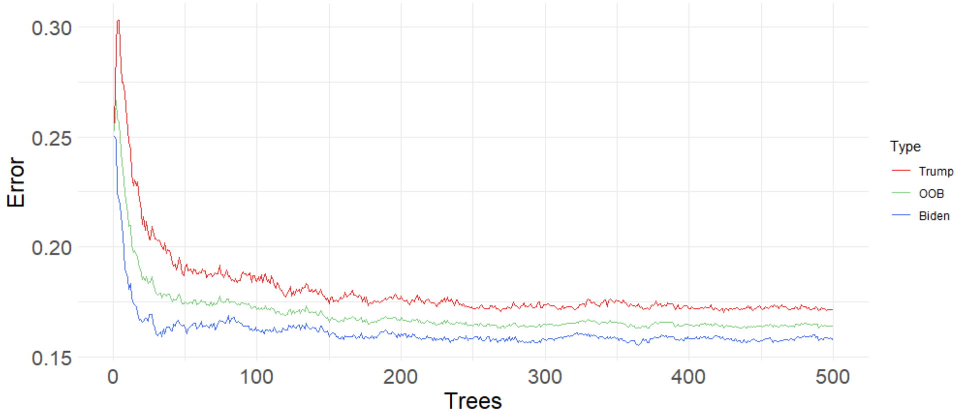


Figure 5.2: Random Forest Error rate against the number of trees used.

Figure 5.2 shows the error values for Trump, Biden and OOB alongside the number of trees we use. The figure highlights the difference in error rates between Trump and Biden, indicating that a Trump vote may be harder to predict than a Biden vote for the variables considered in this analysis. From the figure, 500 trees seem to be a good number as the variance of the individual error rates really dies out around here.

As mentioned in Section 4.2, we use the Mean Decrease in Impurity (MDI) using the Gini Index to rank the importance of our predictor variables. Figure 5.3 below shows the ranking of this importance. The top five predictors here are defence spending, abortion, immigration policy, death penalty and ethnicity. Random Forest ranks four variables in the top five in common with the logistic model, swapping out Liberal/Conservative identity for the death penalty. In particular, the number one predictor here, defence spending, has swapped rankings with the logistic number one predictor, immigration policy. Notably, of the four opinion variables considered in this analysis, three of them take the top three ranks in both models. Another notable difference in the graphs is their shape. The Logistic model has a sharp drop at the death penalty and another sharp drop after religion. Conversely, Random Forests' importance is a much more steady drop off, with the biggest shift seen from first to second. From this, we can say that opinion variables seem to be the most important predictor of vote choice compared to demographics and socioeconomics.

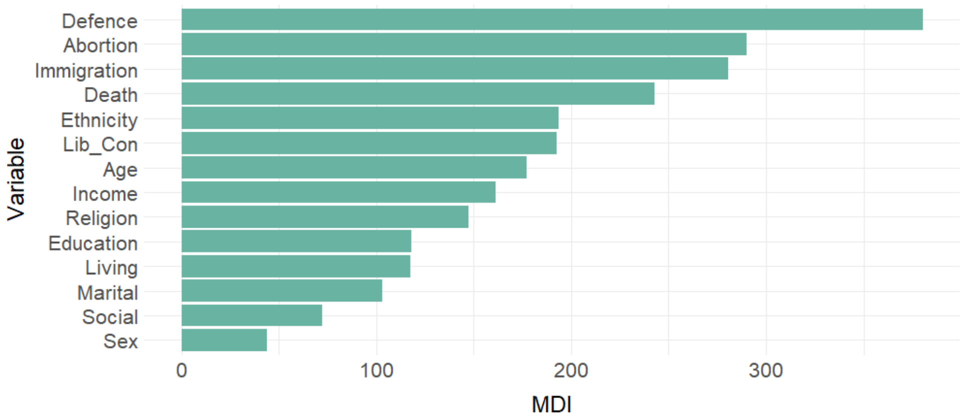


Figure 5.3: Random Forest Importance ranked via MDI using Gini Index.

5.3 Model Comparison

In this section, we will compare the performance of the Multiple Logistic Regression and Random Forest model using different metrics. We have already compared the ranking of importance by these two models so here we will explore other performance metrics. Table 5.2 below compares the two models on some different metrics. Overall, both models perform well with similar F1-scores. The Logistic model has a slightly higher overall accuracy using 10-Fold CV. We can see that both models seem to predict Biden votes (specificity) better than Trump votes (sensitivity) with the Logistic model having the most notable difference here. The Logistic model may be performing better here if the data is proving to be linear, it being designed for linear data whereas Random Forest can accommodate both linear and non-linear relationships.

Model	10-fold CV Acc.	Sensitivity	Specificity	F1-Score
Logistic	0.84	0.82	0.87	0.82
Random Forest	0.82	0.83	0.84	0.81

Table 5.2: Performance metrics for Random Forest.

After evaluating the performance metrics, we examined the ROC curves depicted in Figure 5.4 for both models. Both models demonstrated strong discriminative ability, as evidenced by the AUC score of 0.92. The lines mostly overlap each other, except when the true positive rate is between 0.6 and 0.9, where the Multiple Logistic Regression model has a slightly superior trade-off. This indicates that the Logistic model has a better trade-off with a higher Trump vote accuracy in this region.

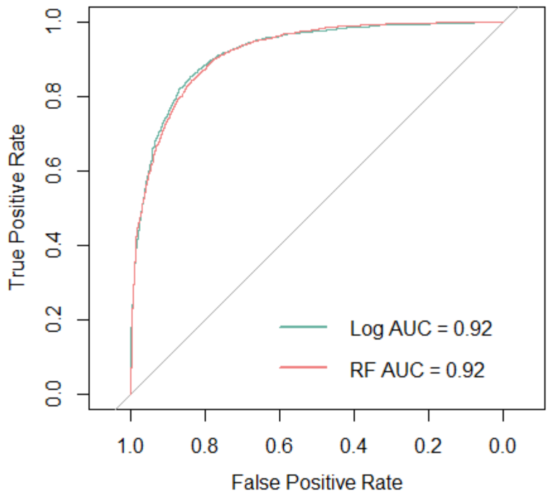


Figure 5.4: ROC curve for both models yielding identical an AUC = 0.92

5.3.1 Evaluating Model Techniques

Multiple Logistic Regression and Random Forest models fit the data to the models in fundamentally different ways. Logistic regression assumes a linear relationship between the dependent variable and the predictor variables. It models the log-odds of the dependent variable as a linear combination of the predictors, making it well suited for capturing linear relationships within the data. Conversely, Random Forest models are highly flexible and capable of capturing complex, non-linear relationships. They do so by constructing an ensemble of decision trees, each of which considers a random subset of predictors at each split, allowing for the modelling of intricate interactions and non-linear patterns in the data.

However, this flexibility in the Random Forest comes with potential drawbacks. Although Random Forests use techniques such as bagging, they are still prone to overfitting. This can make the Random Forest capture noise in the data rather than true underlying patterns. Random Forests are additionally sensitive to various hyperparameters such as the number of variables tried at each split, maximum depth, the total number of trees and so on. The selection of these hyperparameters can significantly impact the model performance.

When it comes to model selection and size reduction, Multiple Logistic Regression and Random Forest models diverge in their approaches. In logistic regression, changes in the AIC can often serve as a reliable guide for removing unnecessary variables and reducing model size. This approach balances goodness of fit with model complexity, leading to a more parsimonious model.

However, with Random Forest models, the process of model selection is less straightforward. AIC can not inherently be used here as Random Forest is not modelled using maximum likelihood and there is no obvious likelihood function for it. From this, it is often very complex to penalise model complexity for Random Forest.

Chapter 6

Discussion, Conclusions and Further Work

This paper has discussed the key demographic, socioeconomic and political opinion variables associated with voting patterns in the US 2020 Presidential Election.

In this comprehensive literature review and exploratory analysis, we have delved into studies that offer intricate insights into the factors influencing voting patterns in past US Presidential elections. The works of Sheng, Garand, Abramowitz and more have contributed significantly to our understanding of demographic, socioeconomic, and political variables impacting voter behaviour. Sheng's examination of the 2016 US Presidential Election highlighted the importance of demographic factors, revealing the nuanced impact of variables such as race, gender, education, partisanship, age, and income. Garand's study, examining immigrant threat and voter choice found that those with an American identity had stronger opinions on immigrant threat and subsequently had a strong effect on vote choice. Abramowitz's focus on abortion opinions in the 1992 US Presidential Election, found abortion to have the strongest influence on candidate choice than any other policy issue considered in the study, from the death penalty to the Gulf War.

Building on these studies, our analysis of the ANES 2020 Time Series dataset aimed to explore the key characteristics of demographic, socioeconomic, and political opinion variables. The dataset, comprising 5729 entries, allowed us to investigate voting patterns based on sex, age, ethnicity, social class, education, income, religion, marital status, Liberal/Conservative identity, living area, abortion, death penalty, defence spending and immigration policy.

Demographically, our analysis unveiled intriguing patterns across various groups, highlighting the significant influence of sex and ethnicity. Males were almost 34% more likely to vote for Trump than females. Additionally, White Americans, constituting a majority of the dataset, demonstrated diverse preferences, with Black Americans overwhelmingly favouring Biden.

Socioeconomically, Conservative identity naturally leaned moderately towards Trump, while Liberals exhibited a stronger preference for Biden. Marital status and religious affiliations also displayed interesting patterns, with never-married participants displaying a clear preference for Biden and varied preferences among religious groups.

Political opinion variables proved to have the most contrasting differences in voter choice. Those opposing legal permission for abortion tended to favour Trump, while those supporting it leaned towards Biden. A similar pattern emerged regarding the death penalty, with Biden receiving more support from those opposed to it. On defence spending, a parallel trend was observed, with a stronger Biden preference among those advocating for decreased spending. With immigration policy, Trump voters tended to have stricter views than Biden voters on the policies for unauthorised immigrants living in the US.

This paper employs a Multiple Logistic Regression and Random Forest model to analyse how demographic, socioeconomic, and political opinion variables predict votes for Biden or Trump. The Logistic model initially includes all variables from the exploratory analysis, later removing income and education due to a weak change in AIC. The model found immigration policy, abortion, defence spending, Liberal/Conservative identity and ethnicity to be the top five most important variables. Noteworthy findings include; those who would like the US govt. to make all unauthorised immigrants felons and send them back to their home country were over 30 times more likely to vote for Trump than those who would like to allow all unauthorised immigrants to remain in the US and eventually qualify for citizenship without penalties, Black Americans were approximately 20 times less likely to vote for Trump than White Americans and those who believe abortion should never be permitted were approximately four and a half times more likely to vote for Trump than those who believed it should be a matter of personal choice. Model evaluation, shows 0.82 sensitivity for Trump and 0.87 specificity for Biden, indicating the model was better at predicting Biden votes than Trump votes. 10-fold Cross-Validation yielded an overall accuracy of 0.84, aligning with the overall accuracy found in the existing model of 0.85, affirming its reliability.

The Random Forest model includes all variables from the exploratory analysis. The model found defence spending, abortion, immigration policy, death penalty and ethnicity to be the top five most important variables. Model evaluation, shows 0.83 sensitivity for Trump and 0.84 specificity for Biden, showing comparable performance on the two candidates. 10-fold Cross-Validation yielded an overall accuracy of 0.82, closely behind the overall accuracy found in the existing model of 0.85, affirming its reliability. The two models perform comparably with the Logistic model proving better at estimating Biden votes than the Random Forest.

Areas of further research have been identified during the course of work carried out for this paper. Notably, further work should aim to look at more opinion variables as the opinion variables in this paper proved, on average, to be the most important in both models. Additionally, it may prove beneficial to create some interaction terms to identify certain types of voters. Other statistical techniques such as different forms of cluster analysis may be useful in identifying different groups of individuals. Additionally, a comparison between the 2020 and 2024 US Presidential Elections may be of interest if the primary candidates of the upcoming election end up being Trump and Biden again.

Bibliography

6.1 References

[1] Campbell, A., 1980. The American voter. University of Chicago Press.

[2] Leighley, J. E., & Nagler, J. (2013). Who votes now? Demographics, issues, inequality, and turnout in the United States. Princeton University Press.

[3] Lupia, A., 2008. Procedural transparency and the credibility of election surveys. Electoral Studies, 27(4), pp.732-739.

[4] Kellstedt, P. M., & Guy Whitten, D. (2017). The fundamentals of political science research. Cambridge University Press.

[5] DeBell , M. (2010) How to analyze anes data - american national election studies. Available at: https://electionstudies.org/wp-content/uploads/2018/05/HowToAnalyzeANESData.pdf (Accessed: 14 October 2023).

[6] Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. Public Opinion Quarterly, 70(5), 646-675.

[7] Lin, F.Y. and Wang, C.H., 2020. Personality and individual attitudes toward vaccination: A nationally representative survey in the United States. BMC public health, 20, pp.1-8.

[8] Murray, G.R. and Scime, A., 2010. Microtargeting and electorate segmentation: data mining the American National Election Studies. Journal of Political Marketing, 9(3), pp.143-166.

[9] Baccini, L., Brodeur, A. and Weymouth, S., 2021. The COVID-19 pandemic and the 2020 US presidential election. Journal of population economics, 34, pp.739-767.

[10] Clem Brooks (2014) Introduction: Voting Behavior and Elections in Context, The Sociological Quarterly, 55:4, 587-595

[11] Aldrich, John H., et al. ”Foreign policy and the electoral connection.” Annu. Rev. Polit. Sci. 9 (2006): 477-502.

[12]Brooks, Clem, Kyle Dodson, and Nikole Hotchkiss. ”National security issues and US presidential elections, 1992–2008.” Social Science Research 39.4 (2010): 518-526.

[13] Bartels, Larry M. “Partisanship and Voting Behavior, 1952-1996.”American Journal of Political Science, vol. 44, no. 1, 2000, pp. 35–50.JSTOR, https://doi.org/10.2307/2669291. Accessed 9 Nov. 2023.

[14] United States: Racial Resentment, Negative Partisanship, and Polarization in Trump's America. Alan Abramowitz and Jennifer McCoy. The ANNALS of the American Academy of Political and Social Science 2018 681:1, 137-156.

[15] Abramowitz, A.I. and Webster, S.W. (2018), Negative Partisanship: Why Americans Dislike Parties But Behave Like Rabid Partisans. Political Psychology, 39: 119-135. https://doi.org/10.1111/pops.1247

[16] Alan I. Abramowitz, Steven Webster, The rise of negative partisanship and the nationalization of U.S. elections in the 21st century, Electoral Studies, Volume 41,2016, Pages 12-22

[17] Iyengar, Shanto, and Masha Krupenkin. ”The strengthening of partisan affect.” Political Psychology 39 (2018): 201-218.

[18] Webster, S.W. Anger and Declining Trust in Government in the American Electorate. Polit Behav 40, 933–964 (2018). https://doi.org/10.1007/s11109-017-9431-7

[19] MORISI, D., JOST, J., & SINGH, V. (2019). An Asymmetrical “President-in-Power” Effect. American Political Science Review, 113(2), 614-620. doi:10.1017/S0003055418000850

[20] Hollibaugh Jr, G.E. (2016) ‘Presidential Appointments and Public Trust’, Presidential studies quarterly, 46(3), 618–639, available: https://doi.org/10.1111/psq.12298

[21] Cary Wu, Rima Wilkes, David C. Wilson; Race & Political Trust: Justice as a Unifying Influence on Political Trust. Daedalus 2022; 151 (4): 177–199. doi: https://doi.org/10.1162/daed_a_01950

[22]Brady, H.E., Verba, S., Schlozman, K.L. (1995) ‘Beyond Ses: A Resource Model of Political Participation’ The American Political Science Review, Jun. 1995, Vol. 89, No. 2 (Jun. 1995), pp. 271-294, available: https://doi.org/10.2307/2082425

[23]Blais. A., Rubenson, D. (2012) ‘The Source of Turnout Decline: New Values or New Contexts?’ Comparative Political Studies, Volume 46, Issue 1, January 2013, Pages 95-117, available: https://doi-org.proxy.lib.ul.ie/10.1177/0010414012453032

[24]Ramakrishnan, S.K., Espenshade, T.J. (2001) ‘Immigrant Incorporation and Political Participation in the United States’ International Migration Review, Vol. 35, Issue 3, available: https://doi.org/10.1111/j.1747-7379.2001.tb00044.x

[25]Argyle, L.P. and Pope, J.C. (2022) ‘DOES POLITICAL PARTICIPATION CONTRIBUTE TO POLARIZATION IN THE UNITED STATES?’ Public Opinion Quarterly, Vol. 86, No. 3, 2022, pp. 697–707, available: https://doi.org/10.1093/poq/nfac036

[26]Yaser, N., Mahsud, N. and Chaudhry, I.A., 2011. Effects of exposure to electronic media political content on voters’ voting behavior. Berkeley Journal of Social Science, 1(4), pp.1-22.

[27]Moy, P. and Hussain, M.M., 2011. Media influences on political trust and engagement. The Oxford handbook of American public opinion and the media, pp.220-250.

[28] Davies, J.J., 2009. The effect of media dependency on voting decisions. Journal of Media Sociology, 1(3/4), pp.160-181

[29] Alsem, K.J., Brakman, S., Hoogduin, L. and Kuper, G., 2008. The impact of newspapers on consumer confidence: does spin bias exist? Applied Economics, 40(5), pp.531-539.

[30] Cohen, J. and Tsfati, Y., 2009. The influence of presumed media influence on strategic voting. Communication Research, 36(3), pp.359-378.

[31] McCombs, M.E. and Shaw, D.L., 1972. The agenda-setting function of mass media. Public opinion quarterly, 36(2), pp.176-187.

[32] Gavin, N.T., 1997. Voting behaviour, the economy and the mass media: Dependency, consonance and priming as a route to theoretical and empirical integration. British Elections & Parties Review, 7(1), pp.127-144.

[33] Weeks, B.E. and Garrett, R.K., 2014. Electoral consequences of political rumors: Motivated reasoning, candidate rumors, and vote choice during the 2008 US presidential election. International Journal of Public Opinion Research, 26(4), pp.401-422.

[34] Garrett, R.K., 2011. Troubling consequences of online political rumoring. Human Communication Research, 37(2), pp.255-274

[35]Biswas, A., Ingle, N. and Roy, M., 2014. Influence of social media on voting behavior. Journal of Power, Politics & Governance, 2(2), pp.127-155.

[36] Intyaswati, D., Maryani, E., Sugiana, D. and Venus, A., 2021. Using media for voting decision among first-time voter college students in West Java, Indonesia. Academic Journal of

Interdisciplinary Studies, 10(1), pp.327-339.

[37] Allcott, H. and Gentzkow, M., 2017. Social media and fake news in the 2016 election. Journal of economic perspectives, 31(2), pp.211-236

[38]Hinds, J., Williams, E.J. and Joinson, A.N., 2020. “It wouldn’t happen to me”: Privacy concerns and perspectives following the Cambridge Analytica scandal. International Journal of Human-Computer Studies, 143, p.102498.

[39] Aday, S., 2010. Leading the charge: Media, elites, and the use of emotion in stimulating rally effects in wartime. Journal of Communication, 60(3), pp.440-465.

[40] Gunther, R., Beck, P.A. and Nisbet, E.C., 2018. Fake news did have a significant impact on the vote in the 2016 election: Original full-length version with methodological appendix. Unpublished manuscript, Ohio State University, Columbus, OH.

[41] Sheng, Y.Y. (2022) ‘The Whos and the Whys Behind Donald Trump’s Victory in the 2016 U.S. Presidential Election’, Malaysian Journal of Social Sciences and Humanities, 2022, Vol. 7, Issue 6, available: https://doi.org/10.47405/mjssh.v7i6.1532

[42] Zingher, J.N. (2020) ‘On the measurement of social class and its role in shaping white vote choice in the 2016 U.S. presidential election’, Electoral Studies, April 2020, Vol. 64, available: https://doi.org/10.1016/j.electstud.2020.102119

[43] Boxell, L. (2020) ‘Demographic change and political polarization in the United States’, Economics Letters, Vol. 192, July 2020, available: https://doi.org/10.1016/j.econlet.2020.109187

[44] Weissberg, H.F. (1987) ‘The Demographics of a New Voting Gap: Marital Differences in American Voting’ Public Opinion Quarterly, Vol. 51, Issue 3, FALL 1987, Pages 335–343, available: https://doi.org/10.1086/269039

[45] Garand, J.C., Qi, D. & Magaña, M. ’Perceptions of Immigrant Threat, American Identity, and Vote Choice in the 2016 U.S. Presidential Election.’ Polit Behav 44, 877–893 (2022). https://doi.org/10.1007/s11109-020-09644-z

[46] Oliver, J. E., & Rahn, W. M. (2016). ’Rise of the Trumpenvolk: Populism in the 2016 Election.’ The ANNALS of the American Academy of Political and Social Science, 667(1), 189-206. https://doi.org/10.1177/0002716216662639

[47] Abramowitz, A.I. (1995) “It’s Abortion, Stupid: Policy Voting in the 1992 Presidential Election,” The Journal of Politics, 57(1), 176–186, available: https://doi.org/10.2307/2960276.

[48] ANES, available: https://electionstudies.org/

[49] ANES 2020 Time Series Study, available: https://electionstudies.org/wp-content/uploads/2021/07/anes_timeseries_2020_userguidecodebook_20210719.pdf

[50] United States Census Bureau, available: https://www.census.gov/quickfacts/fact/table/US/PST045222

[51] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021). 'An Introduction to Statistical Learning with Applications in R', 2nd ed., Springer Cham

[52] Berk, R.A (2008). 'Statistical Learning from a Regression Perspective', New York: Springer Science+Business Media, LLC

[53] Leseur, S. (2023). 'What is a decision tree (parts, types & algorithm examples)', Slickplan, 21 Aug, available: https://slickplan.com/blog/what-is-a-decision-tree [accessed 10 Apr 2024].

[54] Guruprasad (2019). 'Notes on Sensitivity, Specificity, Precision, Recall and F1 score.', Medium, 13 Nov, available: https://medium.com/analytics-vidhya/notes-on-sensitivity-specificity-precision-recall-and-f1-score-e34204d0bb9b [accessed 10 Apr 2024].

[55] Burnham, K.P. and Anderson, D.R (2002). 'Model Selection and Multimodel Inference', 2nd ed., New York: Springer-Verlag Inc.