Predicting the 2025 Canadian Federal Election

STA304 - Assignment 3

Group 44: Johnson Vo, Zhenning Guan, Sahil Patel, Vincent Chu

November 5, 2021

Contents

Introduction

While considered a smaller country in contrast to the many megalomanic powerhouses of the world, it comes to a surprise to many the impact Canada wields on a international scale. It is of great interest then in considering the policies and positions of the country and how it affects all of its stakeholders. Regardless of one's political disposition the result of a national election holds great effect. In order to gauge the political and economic environment of the future, it would be incredibly important to try and forecast Canada's political climate by analyzing voter's preferences. More specifically, forecasting the odds of different parties' performance in future elections, such as the 2025 election. By performing forecasts on the 2025 election, we in turn will shed light on the political workings and environment of the Canada government now and into the future. This information is useful not only on an individual level but also for political parties. For example, if the Liberal party is forecasted to win in a landslide, the conservatives could improve their odds by adjusting their political positions and become more left leaning to acquire more voter support. Furthermore, the outcomes of this and other such forecasts could provide insight towards the winning party into the specific of why they are forecasted to win in order for them to maintain this position for future election [5].

From a behavioral aspect, the forecasted results of an election can influence the behavior of voters [10]. Since one of the aspects of election pits the interest of different socioeconomic groups against one another. For example, if a party supported by a specific demographic group is not forecasted favourably, more individuals within that demographic might get to vote or convince others to participate in the voting process. This will increase turnout rates of votes for that specific party, as the individuals in that demography seek to protect their own interest [10]. Thus forecasting results of future elections is rather important from a political and ethical standpoint, as it increases the information accessed by voters.

In order to have a better comprehension of the Canadian political outlook in the near future, this report will aim to investigate the likely outcome of the 2025 Canadian Federal election using a number of statistical tools. Specifically this forecast aims to determine the winner of the popular vote in the upcoming election. To develop the forecast model for the 2025 Canadian election, two datasets will be used. The first dataset is the General Social Survey conducted in 2017 (GSS2017) by Statistics Canada. This, being a census, is the larger of the two in terms of observations and it will act as representative data of the Canadian population. The second dataset used is the 2019 Canadian Election Study (CES2019) which is the smaller dataset but contains important information on the preference of the voters. The CES2019 dataset will act as the model generating data. Using these two datasets, we've developed a method under a multilevel regression with post-stratification (MRP) approach in order to answer the question of **which party will be the most likely to win the 2025 Canadian federal election.**

As a whole, this model will be built on the highly biased survey data to determine the probability of voters with specific characteristics for a specific party. After the model is generated using the CES2019 data, the GSS2017 data will then be post-stratified and the model will be applied on on it in order to map the results from a the CES2019 dataset to the GSS2019 dataset. This method will allow the model to predict the election outcome with data that more closely resembles the population, providing a more accurate and unbiased forecast using existing data. Although the Canadian Federal Election is determined by seats won in ridings, that data is not available in the census dataset and therefore a popular vote will be used as a proxy, meaning that the model will aim to forecast the results by popular vote. Given the results of the recent election in 2021, it is hypothesized that the conservative party will win the popular vote [6].

Data

As mentioned previously, two datasets will be used as part of this analysis. The CES2019 dataset is a smaller scale survey of citizens conducted each election year, with the mandate to provide detailed information on the election in order to advance the scientific knowledge of voter's motivations as well as the process of democratic elections and their campaigns. Additionally the data displays the political preferences of Canadians to researchers. The CES data for the 2019 election contains a smaller sample of 4,021 observations and detailed information about the background of voters as well as the parties they intended to support. The dataset itself

was collected in two different phases around the 2019 election period by conducting telephone interviews with Canadian citizens using a random digit dialing method to select telephones in order to minimize any survey bias [17]. This dataset was accessed by the cesR library [17].

The GSS2017 dataset is a large dataset focused on familial data of all non-institutionalized persons above the age of 15 that reside in Canada. The two objectives of this survey were to: 1) monitor the changes in the living conditions and well being of Canadians by gathering data on social trends over time; and 2) provide information on current or emerging issues. It was conducted between February 2nd to November 30th in 2017 over telephone interviews over a scope of 39,323 households in which approximately 52.4% (20,602) of households responded [4]. The survey asked questions pertaining to family origins, income, and more. The GSS2017 dataset is a much larger dataset when compared to CES2019 and is will be considered to be representative of the Canadian population in this analysis. However it does not contain information on the political preference of the individuals and thus a model cannot be purely built on it. Post-stratifying the model onto this dataset will allow the model to more accurately forecast the result as GSS2017 more closely resembles the population. This dataset was accessed by the CHASS portal [4].

Cleaning Process

Due to the nature of post-stratification there is a large amount of data cleaning needed to be done in order to be able to run our MRP model. This is performed in the `cleaning_script.rmd` file. We first deal with the survey data (CES 2019 survey). The majority of the dataset consists of different responses to the numerous questions from each person in the survey. Since these are all subjective in nature, the first immediate step in cleaning is to remove most of the data set pertaining to these questions. The remaining variables in the survey dataset are age, gender, citizenship, religion, income, province, education and the most likely vote variable. However in the original dataset these pertain to the question number - age, q3, q1, q62, q69, q61, q4 and q11 respectively. Therefore the next step is to remain the variables to their respective names as stated above.

The citizenship variable is used to verify whether every correspondent in the survey is of Canadian citizenship as only those are actually able to vote in the federal elections. We also need to filter out any N/A values within the dataset. Doing this reduces the survey dataset from 4021 observations to 3937 observations. One notable difference between the survey and the census is that the census data for age contain one decimal digit while the survey age data only contains whole numbers. We round down (i.e floor the values) the census age data values to match the survey age data.

In order for post-stratification to perform properly, we need to rename the values for each of the 6 desired predictor variables in both the census and survey dataset so they match each other. Every variable in the survey is denoted by a number to represent the corresponding answer as denoted in the cesR documentation [17]. Generally, the cleaning process needs to be done on 7 variables in the survey data and 6 variables in the census data - age, gender, citizenship, religion, income, province, education being the 6 shared variables and vote being the 7th in the survey data. The process of cleaning is the same for all 6 shared variables - the first step is we filter out any undesirable values/observations within a variable in the survey data. The second step is using the cesR documentation to translate the number response to a more suitable response using the `mutate()` and `filter()` function in base R. In the third and final step, we mutate the census data so that each of the variable values match the survey values.

For example, the vote variable in the original survey dataset has numerical values -9, -8. -7 and 1-10. Generally in the survey data, negative numerical values are answers such as “Undecided”, “Skipped” or “Refused”, i.e answers that abstain from answering the question or those who are not sure. The first step is to remove individuals with a -9, -8, -7 value as these observations are not particularly useful for our analysis. In particular, -9 corresponds to a “Undecided/Don’t know” answer. While there is merit in retaining this value, since those who are “Undecided” may still end up voting in the election, being unable to ascertain their vote in the actual election will only harm our analysis as this lowers the overall proportion of votes in our model (i.e the model accounts for “Undecided” votes when in the real election you cannot vote “Undecided”). We also remove observation values under 9 and 10, which pertain to “None of these”

and “Will spoil ballet” respectively. Once again these answers to the vote variable do not add any more information for our analysis. Then the second step is to rename the remaining numerical value: 1-7, to their corresponding labels as denoted in the documentation [17]. These labels indicate the intended party the individual wishes to vote for (for example 1 pertains to the Liberal party). Since the vote variable is only contained in the survey data and not in the census data, the third step is not employed here.

This is employed similarly for every other variable with slight differences. For example, the province variable in the survey data does not need to have its values filtered out, since every observation in the dataset are always one of the ten provinces of Canada and not a value such as “Don’t know”. However step 2 and 3 of this process are employed - each numerical answer of the province variable is translated to their corresponding province label as stated in the cesR documentation [17] and the census data is mutated such that each observation has their province value changed to match the survey data.

Another variable to modify is the education variable. Specifically this variable indicates the highest level of education completed. For precision purposes, some of the values in the education variable in the survey data were grouped together as certain groups such as those who have no formal schooling were of low population. This may result in the model being biased on these individuals and underfit on these categories. The original survey had 11 different categories of education. Once again the first step is to remove observations pertaining to “Don’t know” and “Refused” answers. Note that the census data for education has 7 categories total. In order to perform post-stratification properly, we need the variables in both datasets to match. Since the census dataset only describes education beyond a bachelor’s degree as ‘University certificate, diploma or degree above the bachelor’s level” while the survey has both “Master’s degree” and “Professional degree or doctorate”, these two values need to be grouped together in order to post-stratify properly. In the end, we decided to group education into four categories: “No schooling”, “Secondary school diploma”, “College, technical or bachelor’s degree” and “Graduate or professional degree.” The first category, “No schooling” contains the “No schooling”, “Some elementary school”, “Completed elementary school” and “Some secondary/high school” values from the survey data only contains “Less than high school diploma or its equivalent” from the census data. Specifically, these groups in the survey data were grouped together and renamed into the “No schooling” category, while the “Less than high school diploma or its equivalent” was merely renamed into the “No schooling” value. Similarly the “Secondary school diploma” category contains “Completed secondary / high school”, “Some technical, community college, CEGEP, College Classique” and “Some university” in the survey data and the census dataset contains “High school diploma or a high school equivalency certificate”. The “College, technical or bachelor’s degree” category contains “Bachelor’s degree” only in the survey data. In the census dataset, it contains “Trade certificate or diploma”, “Bachelor’s degree (e.g. B.A., B.Sc., LL.B.)”, “College, CEGEP or other non-university certificate or di...” and “University certificate or diploma below the bachelor’s level” categories. Doing this allows for the education variable to match between the two datasets into four concise categories.

The income variable in the survey data is a continuous numerical variable whereas in the census it is a categorical variable. In particular the income variable in the census dataset is named “income_family” and as such in the final step of cleaning this will need to be renamed to income. In order to post-stratify, we need to convert the survey data (which are continuous) values into categories. Therefore the cleaning process for the income variable is to first remove values -9 and -8 which pertain to “Don’t know” and “Refused” values, and then to categorize all of the numerical values in the income variable. These categories match the census dataset family income variable categories: “Less than $25,000”, “$25,000 to $49,999”, “$50,000 to $74,999”, “$75,000 to $99,999”, “$100,000 to $124,999” and finally “$125,000 and more”. This is done by grouping every numerical value in the survey dataset to convert it into a categorical variable. Since we changed the values in the survey dataset directly into the same categories present in the census dataset, this means the census dataset does not need to undergo any changes.

Religion in the context of this analysis is not too focused on the different types of religion and their role on voting habits, but rather on the religiosity itself - i.e we are more concerned over whether an individual is religious rather than what religion they follow. This means for the religion variable in the survey dataset, we can convert the values into a categorical variable with two different responses: “No religious affiliation” and “Has religious affiliation”. The first step is to remove the -8 values in the survey dataset under religion - which pertains to the response of “Refused”, as these responses do not help in our analysis. We consider

those who answers "Don't know/Agnostic" and "None, don't have one / Atheist" under the "No religious affiliation" (whose values are found in the cesR documentation [17]) category and every other response under the "Has religious affiliation" category. Thus transforming the religion variable from 23 different categories to a mere two. On the census side, this means using the religious affiliation variable which contains three categories: "Has religious affiliation", "No religious affiliation" and "Don't know." For the purposes of this analysis, those who answered "Don't know" in the census were grouped together with the "Has religious affiliation" for both brevity purposes and since those who do not know if they are religious themselves tend to be more closely associated with the "agnostic" side than the religious side. Basically those on do not know what religion they follow often are agnostic in nature. This allows both datasets to have their religion variable to match.

Given the perceived benefit of using gender as a predictor and variable in the post-stratification process [11], it will be used as the variable of choice when identifying individuals, as opposed to sex. Given that the gender of the individual is already recorded within the survey data, no form of mutation regarding the gender identity of an individual is needed. However, considering that the methodologies of this study aims to post stratify the census data from models drawn from the survey data to predict the election, gender also needs to be a valid variable for the census data. That being said, biological sex was recorded as the variable of interest for the census data, and thus it must be mutated to gender so that it can be used in the post-stratification stage. This mutation was done by simply mapping all observations who identified their sex as male to their gender being male and similarly for women. Given that this simply one-to-one mapping was done, we must acknowledge the limitations in using equating gender and sex as a predictor.

The predictor age does not need to undergo this process as it already is the same in both datasets after rounding down the census data. However one thing to note in this analysis is that the minimum age of observations in the census data is 15 - below the voting age minimum. However this analysis seeks to predict the 2025 election, and thus at the time of the election the 15 year olds represented in the census will be of age. We concluded that that these individuals will remain the analysis.

After cleaning all 7 variables and removing any N/A values as well as the undesirable "Don't know/Refused" answers, we end up with 2178 observations and 7 variables in the survey and 20051 observations and 6 variables in the census datasets. The final step of cleaning is to ensure the variable names match by renaming the sex, religion affiliation and income family variables into "gender", "religion" and "income" respectively.

Variables

After the cleaning process there are 6 variables left in the GSS2017 dataset and 7 variables left in CES2019 dataset. All 6 variables in GSS2017 also exist in the CES2019 dataset, with the only unique variable, the 7th variable, in CES2019 being the variable vote. The multilevel model will incorporate all 6 variables with age, gender, religion, income, and education on an individual level and province on a group level. The choice of predictor variable is heavily dependent on whether the predictors are capable of explaining variations in the response variable. There is some support that the variables listed above are correlated with political preferences.

Age can play a role when it comes to voting preferences because of generational differences. Each generation typically has a common set of values and beliefs within it. The younger generations might be exposed to different ideas than older generations because of cultural ideological shifts. Therefore, each generation is inclined to be either more left or right wing based on which generation they belong to. Therefore, age will be included in the model and further analysis of the ability of age to explain variation on vote preference will be conducted in the model selection section. Notably the age variable in the census includes 15 year olds. However they are included for this analysis merely on the fact that we are predicting the 2025 Canadian election, and so anyone not of voting age in the census will be of voting age by the time the 2025 election occurs.

Gender will be considered a significant predictor as women historically have been in a more disadvantaged position compared to men. Hence, women are more likely to be sympathetic to those who are in a privileged position in our society as left leaning parties often put more emphasis on helping the disadvantaged and

marginalized groups in our society. In the CES2019 dataset, the difference between male and female voting intention of the Liberal Party has a slight difference of 4.07 %. However, the difference in voting intention for the Conservative Party is larger than 14.32 %.

Females have a significantly higher intention to vote for the Liberal party which also aligns with the findings in Leger's latest federal survey [8].

Income should be a predictor for political preferences since different parties have different political ideologies which ultimately influence the economic policies within the country. One of the most influential economic ideas regarding income is fiscal conservatism. Fiscal conservatism advocates for lower taxes, reduced government spending and minimal government debt. On the other hand, liberal parties advocate for higher tax rates and government spending to combat economic inequality and provide assistance to the economically disadvantaged in society. Since individuals with higher income likely does not need the social assistance provided by the government, the higher tax rate is not beneficial to such earners (in fact it may even be detrimental) which acts as an incentive for them to position themselves more conservative from an economic standpoint. Thus a clean political delineation is made purely through economics and income.

Education can play a role when it comes to voting preferences because people who have post different levels of education might be exposed to different information which shapes the discrepancies in mindset that is divided by their level of education. According to Hare (1981), individuals who are more educated are typically more open minded to new ideas [13]. Hence, voters with a higher levels of education tend to be more progressive and possess a more liberal mindset. This idea is further confirmed by the statistics provided by the Pew Research Center; a report from 2016 states that "Highly educated adults – particularly those who have attended graduate school – are far more likely than those with less education to take predominantly liberal positions across a range of political values"[9].

These conjectures and assumptions can be further elaborate through a number of plots:

Table 1: **Descriptive Statistics of The Data Sets.**

Variable	**Census**, N = 20,051	**Survey**, N = 2,178
AGE	52 (18)	50 (16)
GENDER		
Female	10,915 (54%)	877 (40%)
Male	9,136 (46%)	1,300 (60%)
Other	0 (0%)	1 (<0.1%)
PROVINCE		
Alberta	1,675 (8.4%)	150 (6.9%)
British Columbia	2,441 (12%)	450 (21%)
Manitoba	1,152 (5.7%)	155 (7.1%)
New Brunswick	1,305 (6.5%)	98 (4.5%)
Newfoundland and Labrador	1,070 (5.3%)	98 (4.5%)
Nova Scotia	1,393 (6.9%)	112 (5.1%)
Ontario	5,467 (27%)	460 (21%)
Prince Edward Island	688 (3.4%)	108 (5.0%)
Quebec	3,743 (19%)	392 (18%)
Saskatchewan	1,117 (5.6%)	155 (7.1%)
INCOME		
$100,000 to $ 124,999	2,109 (11%)	268 (12%)
$125,000 and more	4,602 (23%)	669 (31%)
$25,000 to $49,999	4,221 (21%)	300 (14%)
$50,000 to $74,999	3,587 (18%)	396 (18%)
$75,000 to $99,999	2,839 (14%)	294 (13%)
Less than $25,000	2,693 (13%)	251 (12%)
EDUCATION		

Variable	**Census**, N = 20,051	**Survey**, N = 2,178
College, technical or bachelor's degree	10,431 (52%)	1,076 (49%)
Graduate or professional degree	1,814 (9.0%)	378 (17%)
No schooling	3,014 (15%)	109 (5.0%)
Secondary school diploma	4,792 (24%)	615 (28%)
RELIGION		
Has religious affiliation	15,960 (80%)	1,343 (62%)
No religious affiliation	4,091 (20%)	835 (38%)

Table 1 displays the list of shared variables in the two datasets. It depicts the mean of variable age and its standard deviation, as well as the count for the categorical variables and their share (percentage) of the dataset. This table provides insight on the distribution of the survey and census datasets and how they differ. The purpose of using both datasets in this analysis is that the survey data which the model will be on is highly biased due to a number of reasons related to its size and collection process. The purpose of post-stratification is to alleviate this bias by re-weighting the model on a much less biased dataset (which the census surely is). For example, the survey has a 60% proportion of males which clearly is not representative of the population. Other factors include the distribution of the provinces in the survey are not weighted to the true populations which further cements the use of the census dataset. This motivates the idea of using an MRP approach.

Table 2: **Vote distribution for survey data**

	N = 2,178
PARTY	
BQ	87 (4.0%)
CPC	762 (35%)
GPC	228 (10%)
LPC	724 (33%)
NDP	329 (15%)
Other	17 (0.8%)
PPC	31 (1.4%)

Table 2 represents the number of observations for each expected vote found in the CES2019 dataset after undergoing cleaning. It clearly indicates that three parties contain the majority of votes for the 2019 election - the Liberal Party of Canada (LPC), the Conservative Party of Canada (CPC) and the New Democratic Party (NDP). As the goal of this analysis is to predict the popular vote (or rather party with the most votes) in the 2025 Canadian election, it is sufficient to only analyze the three biggest parties as a result. Therefore our model going forth will only consider the three largest Canadian parties. Table 1 and Table 2 were generated with the help of the gtsummary package [16].

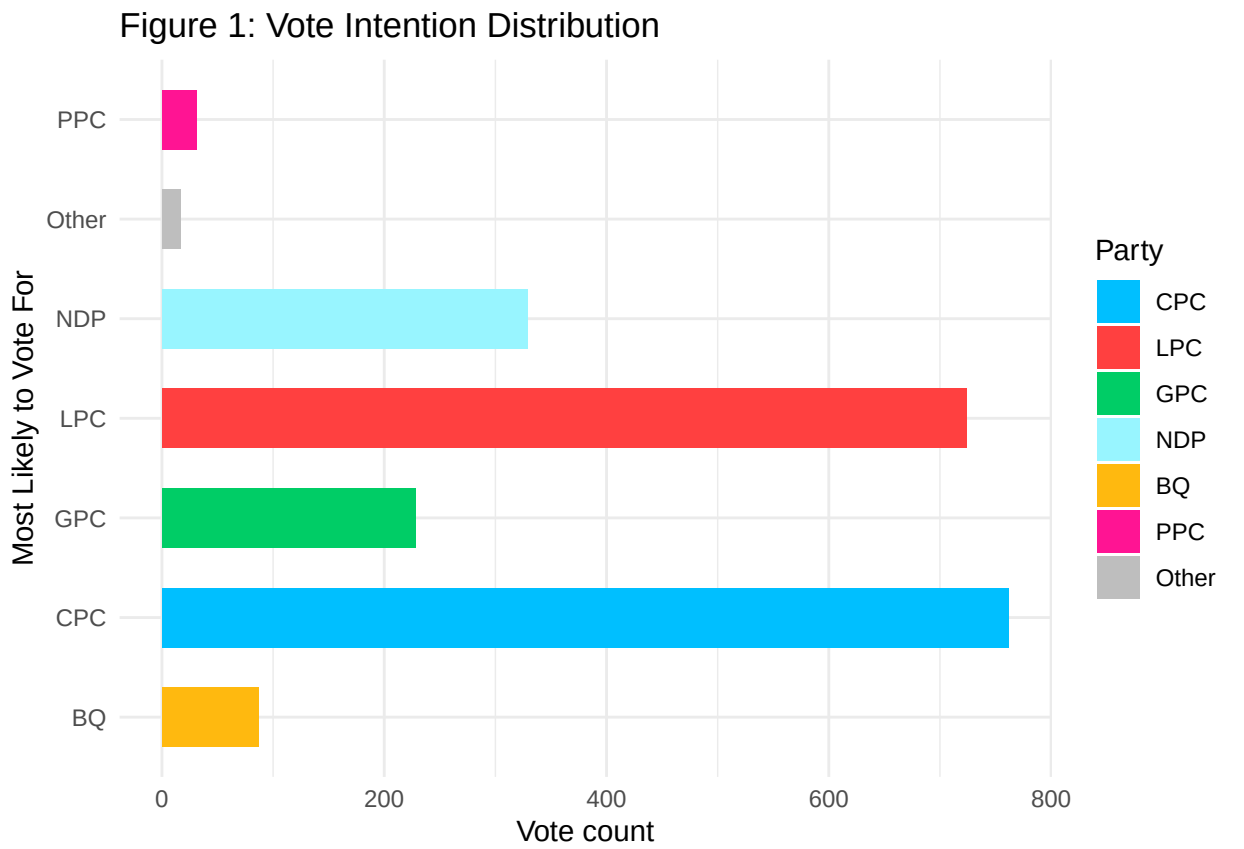
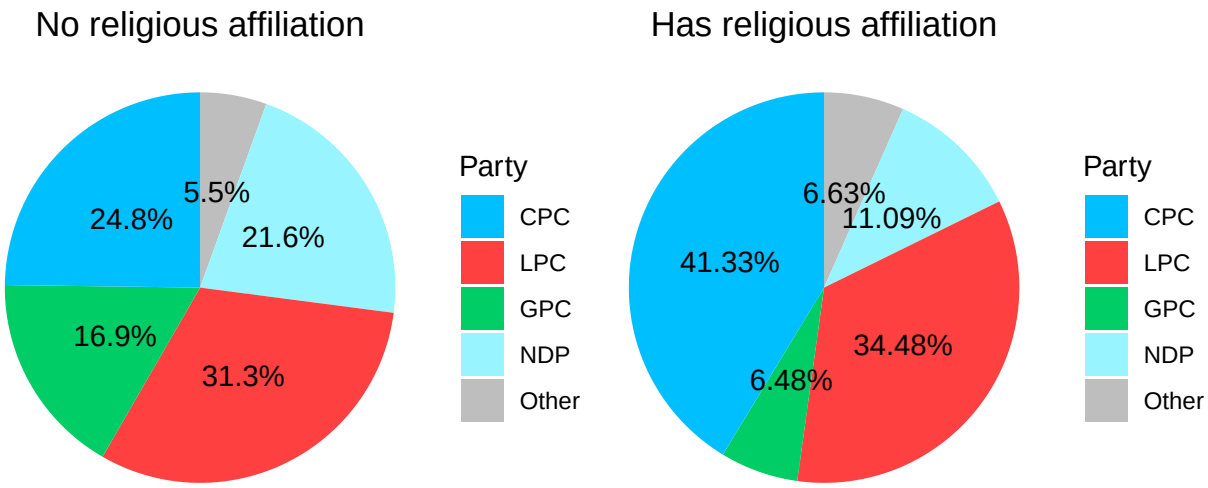


Figure 1 shows the vote intention distribution of participants in the dataset CES 2019, essentially being a visual representation of Table 2. This distribution interestingly maps the 2021 Canadian Federal Election outcome closely [6]. The 6 different Canadian parties are represented by their abbreviations: LPC for the Liberal Party of Canada, CPC for the Conservative Party of Canada, NDP for the New Democratic Party, PPC for the People's Party of Canada, BQ for the Bloc Quebecois, and GPC for the Green Party of Canada. Although the LPC won the overall election (i.e the majority of parliamentary seats) in the 2021 federal election, CPC took the popular vote. In terms of popular vote, LPC comes in second and NDP ranked third.

The results of the federal election depended on seats gained from ridings. Since information regarding ridings is not available in the datasets, the popular vote will be used as a proxy to reflect the public opinion of which party is in favour. According to Figure 1, one of LPC, CPC, and NDP are highly likely to win the popular vote in the coming election. This information led to the selection of parties of interest. The parties selected are LPC, CPC and NDP as these three parties are the most likely to obtain a popular vote, which in turns signals that these three parties are the most likely to be the major players in the next election.

Figure 2: Vote intention for religious vs non−religious groups



A study conducted by Wilkins-Laflamme & Reimer (2019) using CES data from 2004 to 2015 found that the attitudes of voters with religious background are shifting left [20]. Specifically they mention religion has a significant effect on voting attitudes but only for those who vote conservative. While religious conservative voters remain significant in the five elections between 2004 to 2015 as examined in the study, conservative religious voters are now making up a smaller share of the adult population due to their dwindling influence, and as such their influences are fading and their political stance seems to be changing [20]. Figure 2 aims to investigate this relationship of the political preferences of individuals with or without religious background. Specifically to evaluate whether the correlation between religious and Conservative party preferences still stands.

According to the two pie charts shown in Figure 2, there is a clear difference in preference between the two groups. Voters with religious affiliation are more likely to vote for CPC when compared to voters without religious background. While voters with no religious background are more likely to vote for LPC, GDP, and NDP when compared to religious voters. This aligns with the previous findings of Wilkins-Laflamme & Reimer (2019) [20]. More interestingly is how the proportion of Liberal voters remains constant for both religious and non-religious individuals - this motivates the idea that religion plays no role in the voting habits of Liberal voters.

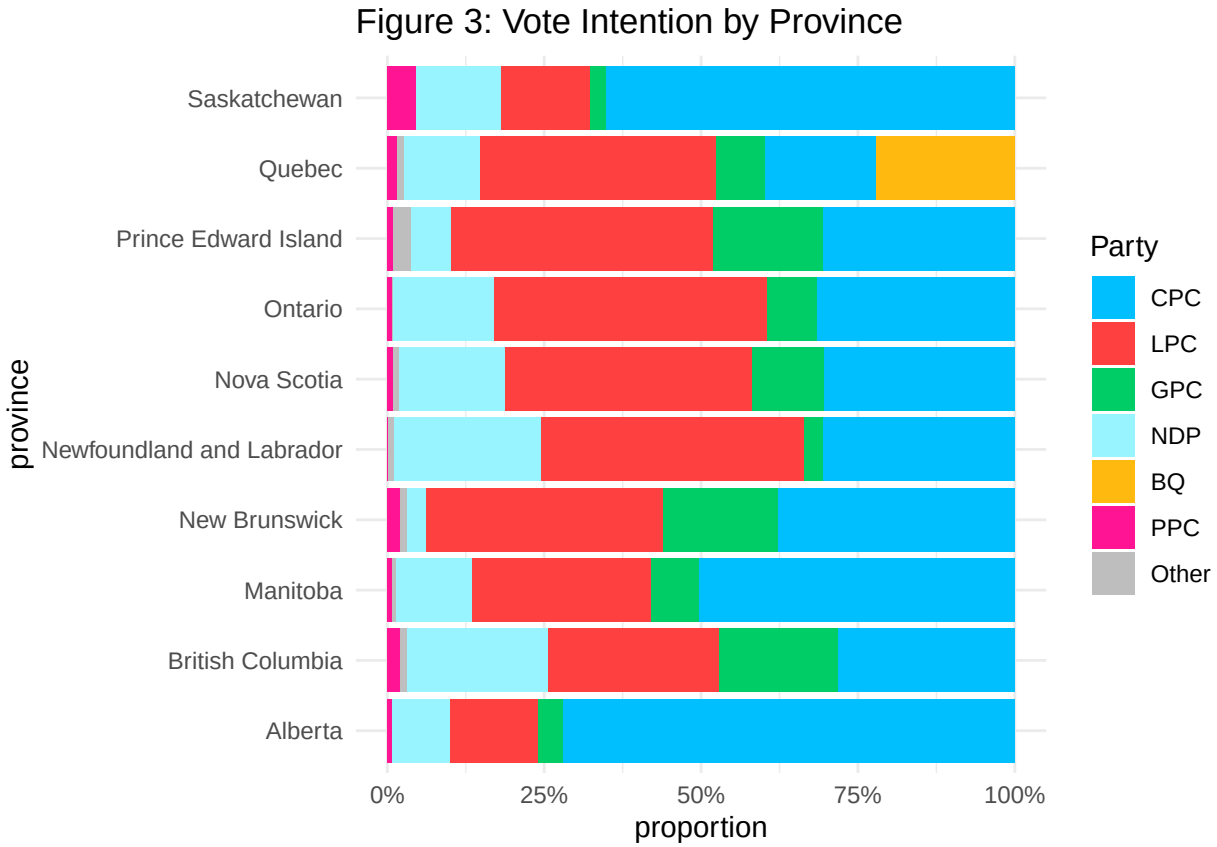


Figure 3 is used to evaluate whether the variable province is suitable to be used as a higher level variable. Since election results heavily depend on the preferences of voters and the location of the voters, choosing province as a higher level variable will allow the model to profile different provinces by cluster. Additionally, it would make sense to group voters by provinces as culture, socio-demographic and economic features differ based on provinces. This is most evidently clear with Quebec, with the Bloc Qubecois party's main purpose to represent Quebec's interests rather.

Figure 3 depicts the ratio of vote intention of the population within each province. There is a clear difference in political preferences in different provinces. While Alberta had almost three quarters of the surveyed respondents having the intention to vote for CPC, only slightly above a quarter of surveyed respondents in Ontario had the intention to vote for CPC.

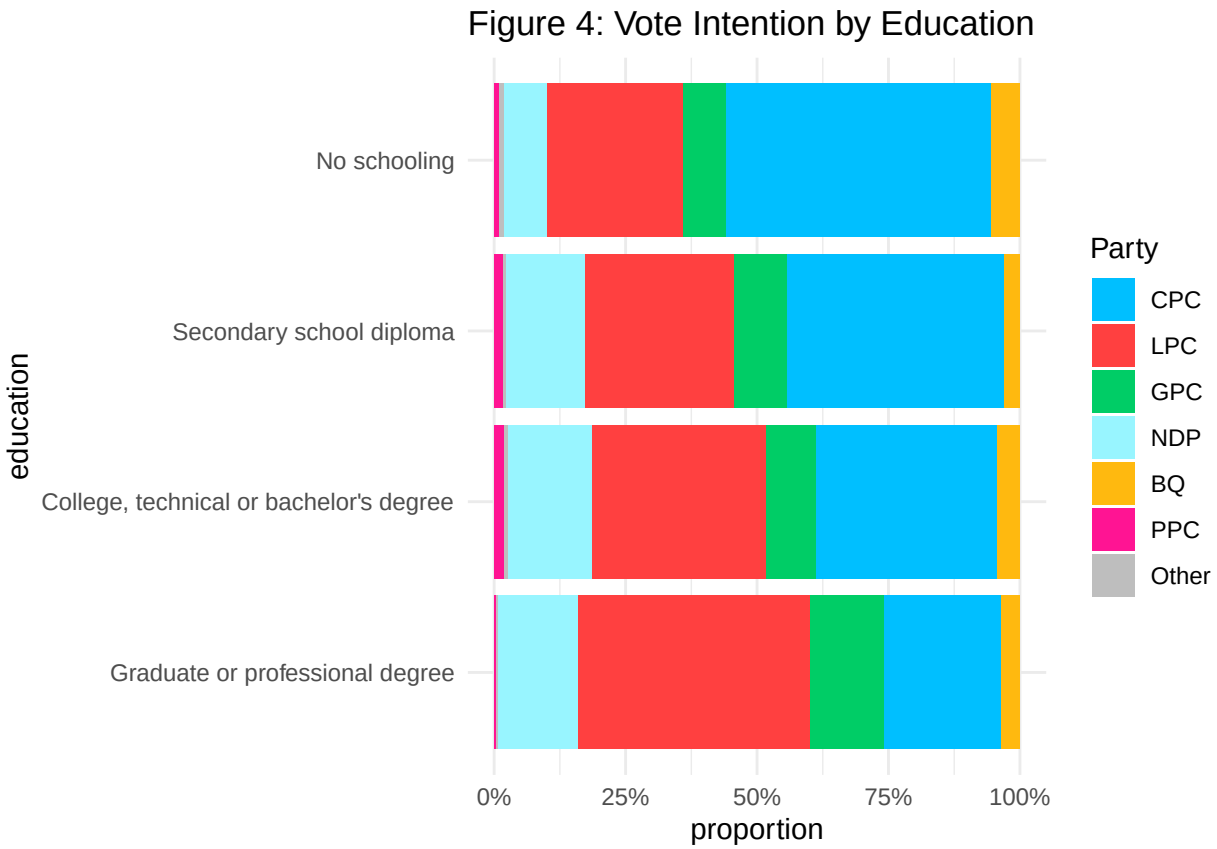


Figure 4 plots the proportion of the voting intention of the survey participants against differing levels of education, which are divided into the four different levels. As stated earlier, previous studies have suggested that higher levels of education tend to have a more progressive and liberal mindset. Figure 4 makes this obviously clear with the increase in the proportion of Liberal voters as the level of education increases. The Figure also represents a decline in the proportion of Conservative voters as hypothesized before. This further reinforces the idea that education has a profound impact on voters preference.

Summary of Variables

Since a multilevel model is being used, the variables listed below are being analyzed on different levels. In this analysis, province is selected to be the second level variable since the election process is heavily dependent on votes from different specific locations. The remaining variables are analyzed on an individual level. After the cleaning process, the variables are represented by the following:

The variable age is a numeric continuous variable with a mean age of 52 years which only takes on discrete values. It has a minimum age of 15 and a maximum age of 80 for the census data. Age has a minimum age of 18 and a maximum age of 95 for the survey data.

The variable gender is a categorical variable and contains three values, male, female, and others. Note that in the census, the variable sex is used, therefore the GSS2017 data only contains two categories for this category, which was transformed into gender during the cleaning process.

Religion is a categorical variable that only contains two categories “Has religious affiliation” or "No religious affiliation".

Income is a categorical variable with categories from “less than $25,000” to “$125,000 and more” with a $25000 increment in interval.

Province is a categorical variable with 10 different categories representing the 10 provinces, excluding territories

Education is categorical with 4 different categories. The four categories being: "College, technical or bachelor's degree ", "Graduate or professional degree", "No schooling ", and "Secondary school diploma".

Vote is a variable that only exists in the CES2019 data, it is the response variable that is used to fit the model. Vote is a categorical variable that represent which party an individual voted for in the 2019 Federal Election, it contains the values: "BQ", "CPC", "GPC", "LPC", "NDP", "Other", and "PPC".

The response categories of interest are the predicted proportions of population within each province that will vote for LPC, CPC, and NDP. These variables will help determine which party is in favour in a specific province. Ideally ridings should be used, as they directly impact the result of the federal election. However, data regarding ridings is not provided in the data which leads to the use of provinces to replace ridings as a predictor.

Methods

Given the structure of the datasets, a MRP model will be used since the response variable of interest only exists in the smaller CES2019 dataset. A model will first be fitted on the CES2019 data to estimate the relationship between the specific characteristics of the voter. Those estimates will then be post-stratified onto the larger GSS2017 dataset to predict the political preference of the voters. Post-stratification is the statistical technique used to reweight or correct a model's estimates based on a known sample population. It does this by partitioning or dividing up the census data into cells or "demographics" based on the model parameters. The model built on the survey data has its estimates corrected by aggregating each of the demographic cell estimates to a population level estimate by weighting the cells by their relative proportion to the population. The post-stratification technique is used because the GSS2017 dataset is a much larger dataset that more closely resembles the population, i.e it is representative of the Canadian population. However, the GSS2017 dataset does not contain the variable vote, which is the variable required to forecast the result of the upcoming election. Unlike the census data, the CES2019 dataset is not representative of the population, as seen from Figure 1 and other earlier observations. By building a multilevel model purely on the survey data, the model fails to capture the full population as a result of the biased survey. Using post-stratification, the model generated from CES2019 can be mapped onto the GSS2017 dataset, which will provide a more accurate and less biased forecast result. This corrects the model's estimates and becomes a powerful model to estimate the election results.

The MRP model will attempt to predict whether an individual will vote for a certain party based on various demographic characteristics that they possess. The individual level inputs for the model are generated from the variables left in the datasets: gender, age, income, amount of completed schooling, and religiousness; while the group level input consists only of the province where the individual resides. The prediction of whether an individual will vote for a certain party is characterized by the logistic model generated by the CES2019 data. For any given party, this model takes the form:

$$logit^{-1}(p_y, i) = \delta_0 + \delta_1 gender_i + \delta_2 income_i + \delta_3 age_i + \delta_4 schooling_i + \delta_5 religiousness_i + \delta_6 province_{p[i]}$$

This generic model classifies whether an individual, i in their given province p will vote for party y where $y \in [CPC, LPC, NDP]$ based on their gender, income, age, highest level of education and whether they are religious. Note δ_0 is the fixed baseline intercept. The remaining estimates, δ_n for $n \in [1, 2, 3, 4, 5, 6]$, represents the relationship between specific characteristics and the log odds of the individual voting for party y. For example, when estimating the log odds for an individual voting for "LPC", δ_1, which is the estimate representing log odds probability associated with a specific gender, is identical for individuals if

their gender is the same. However, a different model would be fitted to estimate the relationship between specific characteristics and the log odds of the individual voting for the conservative party, i.e CPC. Since this model is a multilevel regression model, these variables are estimated as a fixed effect. Furthermore this model varies by province, i.e by the random effects of the second level estimates.

The model shown above is the full model for each party. However to optimize the prediction models, this model will be the base model used for model selection to evaluate the significance of predictors. Note that the models used to estimate different parties might consist of different variables as model selection is performed for each model specific to a party.

For the variable income, it is divided into 6 categories ranging from "$25,000 and below", to "$125,000 and above" with a $25000 increment. δ_2 represents the connection between the log odds of an individual voting for the party of interest, and the income segment that the individual belongs to. The fourth estimate δ_3 represents the relationship between log odds of an individual voting for a specific party and the age of the individual, given that all other variables remain constant. The variable age is discrete, it takes on the rounded integer value of the age of the individual, however it is considered a continuous variable in the context of this analysis. The next estimate in the equation is δ_4 and it represents the relationship between the category of education received and log odds for the individual to vote for the given party. The variable education is categorical, taking on one of the following four values: "College, technical or bachelor's degree", "Graduate or professional degree", "No schooling", or "Secondary school diploma". δ_5 estimates the relationship between the log odds of an individual voting for a specific party and the religious background of the individual. The variable associated with the estimate is categorical, taking on one of two possible values of either "Has religious affiliation", or "No religious affiliation". Lastly, δ_6 is the estimate for a group level input. This estimate evaluates the relationship between the province which the voter resides in and the log odds for the voter to support a specific political party. The variable province is categorical and takes on one of 10 Canadian provinces as its value. Variables are included in the full model as they are considered to be able to explain variation in the response variable, log odds of voting for a specific party, according to preliminary research. However, not all variables included might be relevant as political preference and culture of a population shift continuously. Therefore, a model selection method will be implemented to best optimize the selection of predictor variables for the models specific to each party.

Model Selection

In order to evaluate whether the predictors of each of the models generated for different parties have optimal predictors, likelihood ratio tests (LRTs) are conducted. LRTs work by comparing two models, one complete model, and one reduced model. Specifically the reduced model is merely the full model with one predictor missing. The test determines whether there is a statistical difference in the predictive power in the model. The null hypothesis for LRTs states that the smaller model provides as good a fit for the data as the larger model. Then the alternative hypothesis proposes the reduced model does not provide as good as a fit when compared to the full model. Whether the null is rejected depends on the p-value generated from the LRT. We set the cut-off limit of the p-values to be $\alpha = 0.05$, meaning any p-values below 0.05 results in rejecting the null hypothesis. By rejecting the null hypothesis, we state that the full model adds more predictive power compared to the reduced model. This means that by adding the missing predictor to the reduced model, the full model is a stronger model, and as a result this particular predictor should remain in the final model. The process for the testing then is to implement a LRT for each of the 6 predictors of the full model as labeled above for all three parties: LPC, CPC and NDP. When we fail to reject the null hypothesis for a predictor p, this means the reduced model without that particular predictor p is said to be as good as the full model. This means adding p to the model provides no new information and thus is a meaningless predictor, and so p should be removed from the model. The likelihood ratio test statistic is calculated using the following formula:

$$LRT_{stat} = -2ln\left(\frac{L(model_{reduced})}{L(model_{full})}\right)$$

where L() is the likelihood function and follows a Chi Square distribution [14]. The calculation of the p-value is expressed as, $Pr(LRT_{stat} > \chi^2_{(n)})$, where n is our sample size. This process is then repeated in order to find the optimal model specific to each party. After a model is fitted using all the predictors (the full model) for a specific party, reduced models are then fitted where each model has a different variable deducted from the full model. If the p-value generated from the LRT is less than α, it can be concluded that the reduction of the variable in reduced model is statistically different compared to the full model, i.e the full model is a better predictive model with that predictor included. Hence, the variable being evaluated should not be removed and the full model should be used. If the test concludes a predictor should be removed with high p-values, then we perform LRT testing once again, this time considering the reduced model as the new full model and testing goodness of fit on the remaining variables. This process is repeated until LRT concludes no other variable can be removed. Note that this process does not consider province as one of the intentions of this analysis is to investigate the vote distribution across provinces and so this predictor must remain in every model.

Table 3: Liberal Party Likelihood Ratio Test

Model	LRT p-value
No religion	0.8491673
No income	0.8585596
No education	8.4566017×10^{-4}
No gender	0.2638645
No age	6.4755288×10^{-4}

Table 3 represents the LRT tests performed on the Liberal Party full model expressed above. We clearly see from the p-values that the reduced models built without religion or income or gender are greater than the cutoff value of 0.05. Meanwhile the reduced models built without age or education have very small p-values from the LRT tests. This suggests that both age and education are considered significant predictors and that gender, religion and income are not significant in the scope of predicting the Liberal vote. However we cannot remove all three predictors all at once from this round of LRT tests, as this analysis only tested the models missing one predictor. Therefore we need to continue performing LRTs.

Table 4: Reduced Liberal Model LRT

Model	LRT p-value
No Income	0.8570907
No Religion	0.8284368

Notably from Table 3, we note that models without religion or income have very high p-values while gender only had a moderately large p-value. This suggests that gender may still be a significant predictor. Before we test for gender, we need to confirm whether we can remove **both** religion and income safely. Table 4 represents this. Specifically we fit a reduced model missing both religion and income and performed likelihood ratio tests on models missing only one of the two predictors. For example, for the "No Income" section, the reduced model missing both income and religion was tested against a "full" model which was only missing religion and included income. Notably the p-value for this was very high, beyond the cutoff value. This suggests that we fail to reject the null hypothesis of the LRT, suggesting that income is not a significant predictor considering the full model that was missing only religion still had a high p-value. Similarly results in the "No Religion" section. This means that we can safely remove both predictors from the Liberal party model, whereas the previous section of LRTs suggested we could only remove one.

Table 5: Reduced Liberal Model Without Income And Religion LRT

Model	LRT p-value
No Education	1.3214197×10^{-4}
No Age	4.880315×10^{-4}
No Gender	0.2868469

At this point the Liberal party model only consists of 4 predictors - province, age, gender and education. However from Table 3 we noted that gender has a moderate large p-value. Therefore we need to perform LRT once again to see if we need to remove gender as well. This is noted by Table 5. This time the full model in this LRT is the model consisting of the 4 predictors listed above rather than the original 6. This model is tested against a reduced model consisting of only 3 predictors. Table 5 notes that both education and age are significant predictors due to the low p-values. However gender still has a large p-value above the cutoff value of 0.05. Therefore we state that we can remove gender from the model. Since the only remaining predictors are education and age which have high p-values, we conclude the final model to represent the log-odds probability of the Liberal vote is represented by province, age and education. This follows from our previous deductions on these variables in the Data section - specifically on how education is significant for Liberal voters and that religion is not.

Table 6: Conservative Party Likelihood Ratio Test

Model	LRT p-value
No religion	$2.0494489 \times 10^{-14}$
No income	0.0108428
No education	3.6080987×10^{-8}
No gender	$5.6064266 \times 10^{-11}$
No age	0.1183693

At this point, LRTs should be very familiar and so the rest of the analysis will not go as in detail as the Liberal model. We can now move onto the Conservative model, starting with the full model of 6 predictors. Table 6 corresponds to the results of the LRT. Notably the variables religion, income, education and gender have p-values below the cutoff, meaning these variables are significant in the model for the Conservative party. The only variable that does not pass the LRT is age. As a result we can safely remove age from the conservative party model. Since every other predictor passed the testing, we can safely assume the final model representing the conservative vote consists of predictors religion, income, education, gender and province.

Table 7: New Democratic Party Likelihood Ratio Test

Model	LRT p-value
No religion	1.0014052×10^{-5}
No income	8.4970366×10^{-4}
No education	0.429216
No gender	3.7637783×10^{-7}
No age	$3.0229271 \times 10^{-12}$

For the model of the New Democratic Party, the LRTs for variables religion, income, gender and age generate p-value below the cutoff. Therefore, these variables are kept in the model. However, education has a relatively

high p-value meaning education is not informative and lacks predictive power. Hence the variable education is removed from the model for NDP. Since the other p-values are very low, we conclude that no further reduction is necessary. This means the final model representing the NDP vote is modeled by predictors religion, income, gender and age.

Model Assumptions

Given that a model has been fit, it is necessary to verify that the logistic regression satisfies the necessary assumptions in order to be a valid statistical measure of whether an individual will likely vote for a certain party based on their demographic characteristics. These assumptions are derived from Sheather's "A Modern Approach to Regression with R" [15]. The first assumption required to use a logistic regression is that the response is binary. This assumption is clearly met as an individual is either classified to vote for a party or not, and nothing in between. Next we must assume that the samples were collected independently, which can be verified by reviewing the data collection methods for the GSS2019 dataset [4] and the CES2019 dataset [17]. Thirdly, we must assume that there is not any large multicollinearity within our variables. In the reduced models this can be assumed to be true as we've run LRTs to remove statistically insignificant predictors, thus at least one predictor would be identified as redundant and removed if there was a lot of multicollinearity.

Table 8: Generalized Variation Inflation Factors

Terms	LPC GVIF	CPC GVIF	NDP GVIF
age	1.05563942864751	NA	1.06265188499492
education	NA	1.08820274855221	NA
gender	NA	1.0293963064612	1.02017531954082
income	1.00682494147234	1.09431888728694	1.01191879061758
religion	1.05427219641461	1.0174756096777	1.06836291460789

Furthermore we can check the multicollinearity of the predictors by checking the GVIF values (generalized variance inflation factors) for each model [15]. GVIF values are generalized forms of VIF values which are corrected by the number of degrees of freedom of the predictor variables. These values are used to measure multicollinearity in our data - the higher these values are, the more our predictor variables in each model suffer from multicollinearity. From Table 8 we note that all GVIF values are relatively low, as typically cutoff values for VIF values are at least > 5. This confirms that all three of our models satisfy the multicollinearity assumption.

Lastly, logistic regressions require a large amount of data to be accurate and for our models, we assume to have a large enough amount of data as the models are built on the survey dataset which has 2178 observations.

Model Specifics

Given the results of the model reduction, we can now construct base models for each of the 3 parties.

Liberal Party Model:

$$logit^{-1}(p_i) = \delta_0 + \delta_2 income_i + \delta_3 age_i + \delta_5 religiousness_i + \delta_6 province_{p[i]}$$

Conservative Party Model:

$$logit^{-1}(p_i) = \delta_0 + \delta_1 gender_i + \delta_2 income_i + \delta_4 schooling_i + \delta_5 religiousness_i + \delta_6 province_{p[i]}$$

New Democratic Party Model:

$$logit^{-1}(p_i) = \delta_0 + \delta_1 gender_i + \delta_2 income_i + \delta_3 age_i + \delta_5 religiousness_i + \delta_6 province_{p[i]}$$

Notice that the interpretation for each δ_n is the same as what was given earlier within the methods, and as another clarification: the $\delta's$ between the models do not have to be the same, they are merely denoted with the same underscore to reference their similar interpretation, not value.

Post-Stratification

Currently the model describing the individual level has been proposed and then reduced as a result of LRTs. Given that the goal is to forecast the 2025 Canadian election, the models can be post-stratified to determine the likelihood that a certain party will win the election. Consider the individual-level models. For each party, the model can be seen as a classifier that decides whether certain subgroups will vote for a certain party. For example, within the liberal party model, subgroups are decided based on their income level, age, and religiousness. One example of a subgroup could be religious 39-year-olds who make over $125,000 or another subgroup could be non-religious 25-year-olds who make less than $25,000. For each model, every possible combination of the categories represents a subgroup, $s \in \textit{Subgroups}$ in which the model specifies if they would vote for the party that the model is representative of.

Whether a subgroup votes for the model's party or not can be denoted by θ_s, where $\theta \in \{0, 1\}$, meaning the model either predicts that members of the subgroup vote for the party or don't. To post-stratify these results onto the population, first, the census data is split into provinces, as the province was the decided group-level predictor, then for a specific party, the data is split based on the predictors used for that model. The number of observations within that subgroup N_s is recorded, and used in the following formula to calculate the proportion of the popular vote the province receives from a certain party:

$$\hat{y}^{PS}_{province} = \frac{\Sigma_{s \in Subgroups} N_s \cdot \theta_s}{\Sigma_{s \in Subgroups} N_s}$$

.

Then to predict the proportion of the popular vote a party gets on the national level to calculate the true $\hat{y}^{PS}$, the same process needs to get repeated but instead by summing the results over all the provinces $province \in \textit{All Provinces}$ where All Provinces represents all the provinces in Canada, and $N_{province}$ represents the number of observations in a certain province from the census data. The formula used to find the proportion of the popular vote a party will get nationally, as a result, is:

$$\hat{y}^{PS} = \frac{\sum_{province \in All\ Provinces} N_{province} \cdot \hat{y}^{PS}_{province}}{\sum_{province \in All\ Provinces} N_{province}}$$

In simpler terms for brevity, we can reduce this using the notation of Gelman [18] where

$$\hat{y}^{PS} = \frac{\sum_{i=1}^{I} N_i \hat{y}_i}{\sum_{i=1}^{I} N_i}$$

to represent the post-stratification estimates on a national level, where I is the number of cells total and N_i is the size of the i-th cell and $\hat{y}_i$ is the estimate of the vote probability in cell i. Then the province analogue is merely

$$\hat{y}^{PS}_p = \frac{\sum_{i \in I_p} N_i \hat{y}_i}{\sum_{i \in I_p} N_i}$$

where p is the specific province and I_p are the set of cells that are presentable by the province p. When performing MRP we generate cell-level estimates by averaging over the cells in accordance to the population proportion and weight the values to get our desired results. Specifically this means we generate our cell-level estimates from all possibilities of predictors. Since each model is built on different predictors, this means each

model will also have different number of cells. The Liberal model, built on age (77 different values), province (10 categories), income (6 categories) and religion (2 categories) will have 9240 cells. The Conservative party, being built on education (4 categories), province (10 categories), income (6 categories), gender (3 categories) and religion (2 categories) will only have 1440 cells. Finally the New Democratic party model is built on age (77 different values), province (10 categories), income (6 categories), gender (3 categories) and religion (2 categories) and thus will have 27720 different cells. However many of these cells will be empty due to some variables not being possible (specifically genders that pertain to "Other", since the census data has gender as a binary variable). Thus once the cells are all generated, the model is used to estimate $\hat{y}$ values for every cell and are aggregated up to a population level by reweighing the estimate values by the cell's relative proportion in the population. All analysis for this report was programmed using `R version 4.0.2` and models built using the package lme4 [3].

Results

The following tables represent the fixed effects of the multilevel models for all three parties.

Table 9: Liberal Party Model Fixed Effects Coefficients

Terms	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.7613357	0.3161307	-2.4082939	0.0160273
Age	-0.0277088	0.0040121	-6.9063740	0.0000000
$125,000 and more	-0.5301014	0.2126371	-2.4929866	0.0126674
$25,000 to 49,999	0.2682055	0.2303396	1.1643913	0.2442655
$50,000 to 74,999	0.0425903	0.2237851	0.1903181	0.8490599
$75,000 to 99,999	0.0052193	0.2373526	0.0219898	0.9824561
Less than $25,000	0.1926674	0.2432323	0.7921127	0.4282950
No religious affiliation	0.5402422	0.1296779	4.1660298	0.0000310

In Table 9, the coefficients of the individual-level logistic regression model that predicts whether an individual will vote for the Liberal Party according to the various fixed demographic features are presented. All of them hold a similar interpretation to how they were presented within the methodologies. However, it is important to note that the intercept (δ_0) now represents the log-odds for religious individuals who have an income between $100,000 and $124,999, independent of their age. As a result, the other δ's represent categorical deviations away from this subgroup when trying to determine the log-odds of whether an individual will vote for the Liberal Party.

Table 10: Conversative Party Model Fixed Effects Coefficients

Terms	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.6509829	0.2836705	-2.2948562	0.0217414
Is Male	0.7070070	0.1047698	6.7481951	0.0000000
Is Non-binary	-11.9441790	924.7738326	-0.0129158	0.9896950
No religious affiliation	-0.8723212	0.1080969	-8.0698057	0.0000000
$125,000 and more	0.3109920	0.1643852	1.8918485	0.0585112
$25,000 to 49,999	-0.1553757	0.1959897	-0.7927748	0.4279090
$50,000 to 74,999	-0.0771402	0.1818250	-0.4242554	0.6713796
$75,000 to 99,999	-0.0781495	0.1960114	-0.3986989	0.6901151
Less than $25,000	-0.1025757	0.2063781	-0.4970283	0.6191691
Graduate or professional degree	-0.6418921	0.1516126	-4.2337660	0.0000230
No schooling	0.5852199	0.2241073	2.6113384	0.0090189
Secondary school diploma	0.2223474	0.1145204	1.9415532	0.0521912

Similarly to Table 9, Table 10 presents the coefficients for the individual level logistic regression model but for the Conservative Party. It is important to note that δ_0 within this model represents log-odds of an individual voting for the conservative party given that they are a religious female who has an income between $100,000 and $124,999, and has college/technical/bachelor's degree.

Table 11: New Democratic Party Model Fixed Effects Coefficients

Terms	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3432427	0.3322368	-1.0331266	0.3015447
Is Male	-0.6895025	0.1267983	-5.4377875	0.0000001

Terms	Estimate	Std. Error	z value	Pr(>|z|)
Is Non-binary	15.0389730	1916.2170185	0.0078483	0.9937381
Age	-0.0281714	0.0040944	-6.8805233	0.0000000
$125,000 and more	-0.5522078	0.2148664	-2.5700055	0.0101697
$25,000 to 49,999	0.2080743	0.2322380	0.8959528	0.3702780
$50,000 to 74,999	-0.0445401	0.2269445	-0.1962600	0.8444066
$75,000 to 99,999	-0.0575400	0.2400072	-0.2397429	0.8105296
Less than $25,000	0.1171294	0.2463084	0.4755396	0.6344024
No religious affiliation	0.5969985	0.1320741	4.5201773	0.0000062

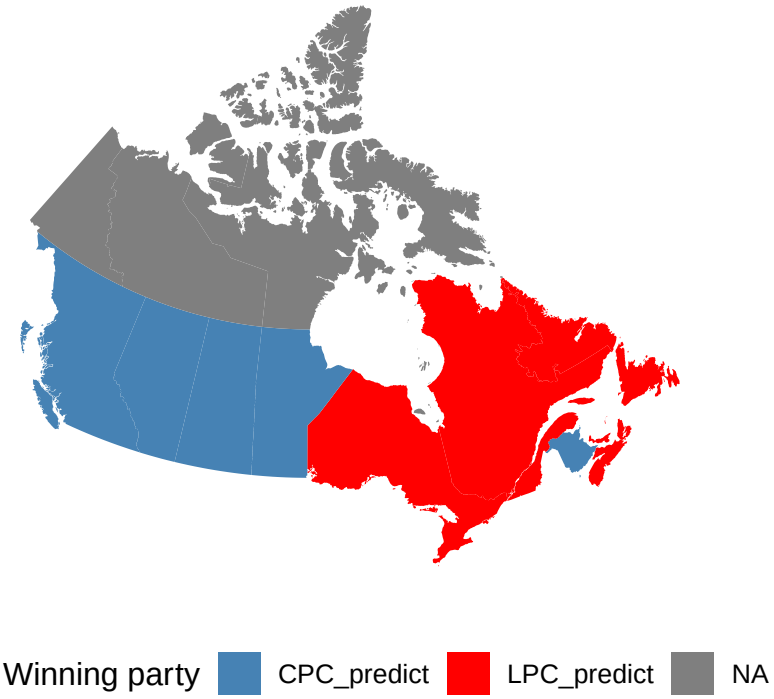
Lastly, Table 11 presents the coefficients for the individual level logistic regression model but for the New Democratic Party. In this model, as opposed to the definition spelled out in the methodologies, δ_0 represents the log-odds of an individual voting for the New Democratic Party given that they are a religious female who has an income between $100,000 and $124,999 independent of their age.

Table 12: Model Random Effects Coefficients

Province	LPC Estimates of Second Level	CPC Estimates of Second Level	NDP Estimates of Second Level
Alberta	-0.3680413	1.3677194	-0.3717460
British Columbia	0.6274931	-0.3368215	0.6286534
Manitoba	0.0197458	0.4224773	0.0610277
New Brunswick	-0.8298975	0.0243309	-0.8867007
Newfoundland and Labrador	0.7173965	-0.4536426	0.7820133
Nova Scotia	0.2168907	-0.3187773	0.2306359
Ontario	0.3589099	-0.3368998	0.3514237
Prince Edward Island	-0.4924417	-0.3315879	-0.5351061
Quebec	-0.1656734	-1.0566984	-0.1775722
Saskatchewan	0.0501572	1.0431838	0.0564072

Table 12 represents the provincial influence on each of the models. These estimates add a level of variance in the log-odds depending on what province the individual is in. Initially the models are fit such that they are averaged over all provinces, however by introducing the provincial estimated coefficients, we notice how the log-odds of voting a certain party might deviate. For example, if the individual is New Brunswick, we suspect a larger decrease in the log-odds when inspecting whether an individual will vote for the Liberal Party.

Figure 5: 2021 Federal Electoral Predictions



Using mapcan[1] we can visualize the results of our analysis. Specifically we take the projected winner of the 2025 election for every province and mapped it out, as seen in Figure 5. Since the survey data contains no information on the territories (Yukon, Northeast Territories and Nunavut) in terms of voter share, the model cannot account for these states. However in an actual election, the results of these three territories will not have a significant result due to only representing 1 seat each out of the total 338 electoral district seats [6]. We clearly see from Figure 5 that the eastern provinces have a larger CPC vote share while the western provinces have a larger LPC vote share and no province has a larger NDP share. Specifically this means that British Columbia, Alberta, Saskatchewan and Manitoba have a higher projected proportion of Conservative voters while Ontario, Quebec, Newfoundland and Labrador, Nova Scotia and Prince Edward Island have a higher projected proportion of Liberal voters. One exception to this region divide is New Brunswick which is projected to have a higher projected proportion of Conservative voters. Note that this analysis is not fully accurate. Firstly, due to only considering the largest parties in terms of votes, the model neglects smaller parties. The Bloc Quebecois is a very popular party in Quebec alone, and in the 2021 election won the majority of seats in Quebec, which this analysis neglects to account for [6]. However as the main purpose was to determine the popular vote winner, we do not necessarily need to consider this issue.

Table 13: Overall Projected Popular Vote

Party	Predicted Voter Share
Liberal Party	33.8%
Conservative Party	36.8%
New Democratic Party	14.7%

As seen in Table 13 we clearly see the projected popular vote winner in the 2025 election by the predicted vote share of the population. Specifically by employing only three models of the largest parties we are able

to pinpoint the projected winner, which is determined to be the the Conservative party, followed closely by the Liberal Party. With this we have predicted the Conservative party will win the popular vote in the 2025 Canadian election.

Conclusions

The goal of this paper was to answer the question of which party will be the most likely to win the 2025 Canadian federal election. Although the critical information of ridings was not provided in the data, the forecasting method aims to use the popular vote as a proxy. The forecast of which party is the most likely going to win the popular vote is achieved by using a multilevel logistic regression with post-stratification and model selection for each party. The model fitted onto the CES2019 data is:

$$logit^{-1}(p_y, i) = \delta_0 + \delta_1 gender_i + \delta_2 income_i + \delta_3 age_i + \delta_4 schooling_i + \delta_5 religiousness_i + \delta_6 province_{p[i]}$$

Where δ_0 is the fixed baseline intercept. $\delta_{1,2,3,4,5}$ are the estimates for individual level variables; gender, income, age, schooling, and religion, and δ_6 is the estimate for the group level variable province. This is the full model.

The significance of the predictor variables in the full models from the dataset CES2019 are then evaluated with a model selection method called likelihood ratio test. Likelihood ratio test assesses the goodness of fit of two competing statistical models. It is expressed as:

$$LRT_{stat} = -2ln\left(\frac{L(model_{reduced})}{L(model_{full})}\right)$$

In the report it is used to compare the predictive power of the reduced model and the full model. If the reduced model can explain as much variation in voters preference as the full model, the predictor that does not exist in the reduced model is then removed due to its lack of predictive power.

After using model selection with LRT for the logistic regression model, it was found that voters' preferences do not depend on the same set for predictors for different parties. The final model for the Liberal Party of Canada does not have the variable gender and schooling as significant; the final model for the Conservative Party of Canada does not include the variable age; and the final model for the New Democratic Party does not consist of the variable schooling.

Liberal Party Model:

$$logit^{-1}(p_i) = \delta_0 + \delta_2 income_i + \delta_3 age_i + \delta_5 religiousness_i + \delta_6 province_{p[i]}$$

Conservative Party Model:

$$logit^{-1}(p_i) = \delta_0 + \delta_1 gender_i + \delta_2 income_i + \delta_4 schooling_i + \delta_5 religiousness_i + \delta_6 province_{p[i]}$$

New Democratic Party Model:

$$logit^{-1}(p_i) = \delta_0 + \delta_1 gender_i + \delta_2 income_i + \delta_3 age_i + \delta_5 religiousness_i + \delta_6 province_{p[i]}$$

These models which are reduced specifically for certain parties are then post-stratified onto the GSS2017 dataset. Since the GSS2017 dataset is much larger than the CES2019 data and by being a census it is representative of the true population distribution. We can aggregate our model predicted values to reduce bias as given:

$$\hat{y}^{PS}_{province} = \frac{\Sigma_{s \in Subgroups} N_s \cdot \theta_s}{\Sigma_{s \in Subgroups} N_s}$$

where θ_s is the estimate of the vote probability in the cell s and N_s is the size of s-th cell.

At the beginning of the report, it was hypothesized that the Conservative Party would win the popular vote according to the data from GSS2017 and CES2019. This was later confirmed with the post-stratified data and subsequent models. According to Table 13, the model fitted onto post-stratified data suggested that CPC will win the popular vote with 36.8% of the population having the intention of voting for the party. Intention to vote for LPC comes in second with 33.8% and NDP is trailing behind with 14.7%.

Another interesting finding is regarding the significance of predictors. The likelihood ratio test implemented in this report as a method of model selection suggests that the significant predictors are different depending on the party of interest when estimating voter's preference. These predictors also have different quantitative values and sometimes even have opposite relationship in terms of vote for specific parties. This is reasonable as left wing and right wing values are the opposite of each other. Comparing the common predictor variables, the most notable and interesting one is regarding income and religious affiliation. According to Table 9 and Table 10, voters with an annual income of $100,000 or less a year are more likely to vote for LPC, while the opposite is true with CPC as voters with annual income of less than $100,000 are less likely to vote for CPC. Religious affiliation is another predictor variable that is not shared among the LPC and CPC model. Similar to income, it exhibited an opposite relationship to voters preference when comparing the variable in the LPC and CPC model. Individuals with no religious affiliation are more likely to vote for LPC and less likely to vote for CPC.

Weaknesses

Despite the results, there does exist a number of drawbacks on the applicability of the results. The first drawback is derived from the sex to gender mapping that was done to the census data so that it could be post-stratified with the survey data. By simply mapping sex to gender, the post stratified results do not take into account the results of any individuals who might not identify their gender as being neither male or female. Furthermore, this mapping assumes that all participant of the survey identify their gender as their birth sex. This method conflates gender identifying males and trans-females into the same ground in terms of sex(and the same for gender identifying females and trans-males), creating a limitation on the interpretability of gender, and as a result, the model as a whole.

Further limitations in the results of the forecasts include the fact that the data used to generate the forecasts were from as recent as 2019. Given that the goal is to forecast the 2025 election, the data, model, and post-stratification will be at least 6 years old when the actual election takes place. Given the constantly shifting political climate, it is possible that the forecasts will be completely wrong. Thus the performance of the forecast is limited by the age of the data used to generate it.

Given the limitations of the gender/sex mapping, a natural next step would be to post stratify using a census dataset that records the gender of the individual. This would remove any gender conflation, and as a result, make the forecast more accurate. Furthermore, since the data will be 6 years old by the time of the 2025 Canadian Federal Election, updating the models using the same procedure and post stratifying onto newly collected census data will provide more accurate forecasts of the election. Another step would be to include factors outside the survey data. For example, we can consider incumbency in the model, i.e candidates who have held the Prime Minister office prior. For example, Justin Trudeau won the 2021 Canadian election and had held the office for a previous two terms at the moment. The act of having experience in office as well as the publicity of being the Prime Minister prior is a considerable factor which may be investigated in a further analysis to improve the model. Another factor outside the datasets may be considering past election results - specifically vote shares in a specific riding and province. This is very important in swing regions - i.e heavily contested ridings that can easiyl flip from one party to another. By considering the previous election results - specifically the vote share of a riding, we can account for these swing regions that are heaviyl contested and unpredictable.

Given that the goal was to forecast the 2025 Canadian election, multilevel modeling and post stratification techniques were used to categorize census data and predict the proportion that certain individuals would vote for various political parties. It was then concluded that the CPC will win the popular vote with 36.8% of the population having the intention of voting for the party. Intention to vote for LPC comes in second

with 33.8% and NDP is trailing behind with 14.7%. However, there are few limitations in the model that originates from the dataset used in this report, which would be improved by using a dataset that contains more minute information about the population, such as the riding of the survey participant and the use of gender instead of sex.

Bibliography

[1] Andrew M, Aaron E (2019). Mapcan: Tools for plotting canadian choropleth maps and choropleth alternatives. R package version 2.1.0. https://cran.r-project.org/web/packages/mapcan/index.html

[2] Baptiste Auguie (2015). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.0.0. http://CRAN.R-project.org/package=gridExtra

[3] Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using lme4." Journal of Statistical Software, 67(1), 1–48. doi: 10.18637/jss.v067.i01.

[4] CHASS. (2020). "General social survey on Family (cycle 31), 2017:"

[5] Daniel Walther. Picking the winner(s): Forecasting elections in multiparty systems, Electoral Studies, Volume 40, 2015. Pages 1-13, ISSN 0261-3794,

[6] Election 2021: CTV News: Canada Election Coverage. (n.d.). Retrieved from https://www.ctvnews.ca/politics/federal-election-2021

[7] Fox J, Weisberg S (2019). An R Companion to Applied Regression, Third edition. Sage, Thousand Oaks CA. Retrieved from https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

[8] Hatemi, P. K., McDermott, R., Bailey, J. M., & Martin, N. G. (2012). The Different Effects of Gender and Sex on Vote Choice. Political Research Quarterly, 65(1), 76–92. http://www.jstor.org/stable/23209561

[9] Ideological Gap Widens Between More, Less Educated Adults. (2020). Retrieved from https://www.pewresearch.org/politics/2016/04/26/a-wider-ideological-gap-between-more-and-less-educated-adults/

[10] Kaplan, B.. (2019). Experimental Electionomics: How Election Forecasts Influence Voter Turnout. Carnegie Mellon University.

[11] Kennedy, Lauren & Khanna, Katharine & Simpson, Daniel & Gelman, Andrew. (2020). Using sex and gender in survey adjustment.

[12] Lumley T (2020). "survey: analysis of complex survey samples." R package version 4.0.

[13] O'Leary, P. T. (1981). Open-Mindedness and Education [Review of Open-Mindedness and Education, by W. Hare]. The Journal of Educational Thought (JET) / Revue de La Pensée Éducative, 15(1), 82–85. http://www.jstor.org/stable/23768244

[14] Philippe J. Fournier June 28, 2. (2020). The biggest divide in Canadian politics? Men vs. Women. Retrieved from https://www.macleans.ca/politics/ottawa/the-biggest-divide-in-canadian-politics-men-vs-women/

[15] Sheather J. S. (2009). A Modern Approach to Regression with R. Springer

[16] Sjoberg, Daniel D., Michael Curry, Margie Hannum, Joseph Larmarange, Karissa Whiting, and Emily C. Zabor. (2021). Gtsummary: Presentation-Ready Data Summary and Analytic Result Tables. https://CRAN.R-project.org/package=gtsummary.

[17] Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, (2020). "2019 Canadian Election Study - Phone Survey", https://doi.org/10.7910/DVN/8RHLG1, Harvard Dataverse, V1

[18] Wang, Wei & Rothschild, David & Goel, Sharad & Gelman, Andrew. (2014). Forecasting Election With Non-Representative Polls. International Journal of Forecasting. 31. 10.1016/j.ijforecast.2014.06.001.

[19] Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686.

[20] Wilkins-Laflamme, S., & Reimer, S. 2019. Religion and Grassroots Social Conservatism in Canada. Canadian Journal of Political Science, 52(4), 865-881. doi:10.1017/S0008423919000544.