Preditive Modeling of Favoured Poltical Parties: A forecast for the upcoming Canadian Fedral Elections

STA304 - Fall 2023 -Assignment 2

Group number : 91, Arnav Dey, Aarnav Chudasama, Shiyan Ahmed Khandaker

11/23/23

Introduction

The objective of this report is to predict the overall popular vote of the NDP, Liberal, and Conservative parties for the upcoming Canadian federal elections in 2025. The report uses data collected from the Canadian Election Study survey conducted in 2021. The data consists of various demographic variables like age, gender, province of its respondents. Moreover, it also contains the political preferences of its respondents by asking them which party they would vote for in the election. However, the collected data is not representative of the total population which necessitates the use of post-stratification on the data. This will make our estimates of the popular vote more representative of the population.

This seeks to tackle an important problem which is the lack of information on the political preferences of Canadian citizens. For example, maybe people of a certain gender or educational standing vote in a specific way that is significantly different from other individuals who are different genders or have a different level of education. This data is useful in order to understand why citizens choose certain parties over others which is important in order to facilitate discussion and dialogue on policies. Additionally, our analysis is important as it allows people, firms and other players in the market and government to prepare for potential political outcomes in the elections.

This analysis is important globally because it gives us an understanding of the changing political landscape between conservative and liberal parties, illustrating changing trends in voter preferences. Furthermore, these political trends give insight into the future of economic and social policies relating to immigration, health, climate etc which will impact the world for decades to come.

This report will include a short description of the data and an explanation of the method by which a logistic regression model was built to predict the popular vote. Then we will summarize our model and poststratification results and provide an analysis of its significance. Based on our findings, we hypothesize that the most popular party will be the Conservative party based on their rising popularity among voters in 2023.

Data

The data we uses were the census data collected by the General Social Survey of 18 variables and 20,602 observations and the survey data collected for the 2021 Canadian Election Study of 20,968 observations of 1,062 variables.

Cleaning:

For the survey data, we use the 2021 CES codebook to implement the provinces by name from the cps21_province column and vote choices by liberal, conservative or NDP from the cps21_votechoice column. From the cps21_genderid, we match the genders numbered from the codebook and remove others and non-binary as we had no data in the census data. We mutate household income, cps21_income_number, to ranges and mutate the education, cps21_education, observations to fit the census data. Using cps21_language_1 which was English and cps21_language_2 which was French, we mutated together to make a language column on whether they speak English, French, both or neither. For both the census and the survey data we modified the number of children to state whether they have children. We rounded the age in the census to only have whole number observations.

Variables:

As predictor variables we pick age, province, sex, language_knowledge, marital_status, children and education. Different age groups often have different priorities, beliefs, and concerns. Provinces may have unique economic conditions, cultural identities, or historical contexts that influence allegiances to certain parties.Men, women and other genders might have different priorities and perspectives on issues such as healthcare, rights, equality and security.Language can be tied to cultural identity and can play a significant role in identity politics.Married people may prioritize family-related policies such as taxation, childcare or education, while unmarried or divorced individuals might focus more on individual rights, welfare and policies. Parents might prioritize education, healthcare and family-oriented policies, while those without children might prioritize other issues such as economic stability, environmental policies, or individual rights. Higher education levels might lead to different perspectives and interests in policies related to healthcare, economy, environment, or social justice.

Sex to gender justification:

Here we have chosen to remove 'non-binary' and 'other' from the gender variable in the survey data. This is in order to ensure survey data matches with the sex variable in census data. This change does prevent us from understanding the voting patterns of non-binary people and those identifying as other genders. However, since this is not a very large number in the survey data, it should not have a disproportionate impact on the analysis.

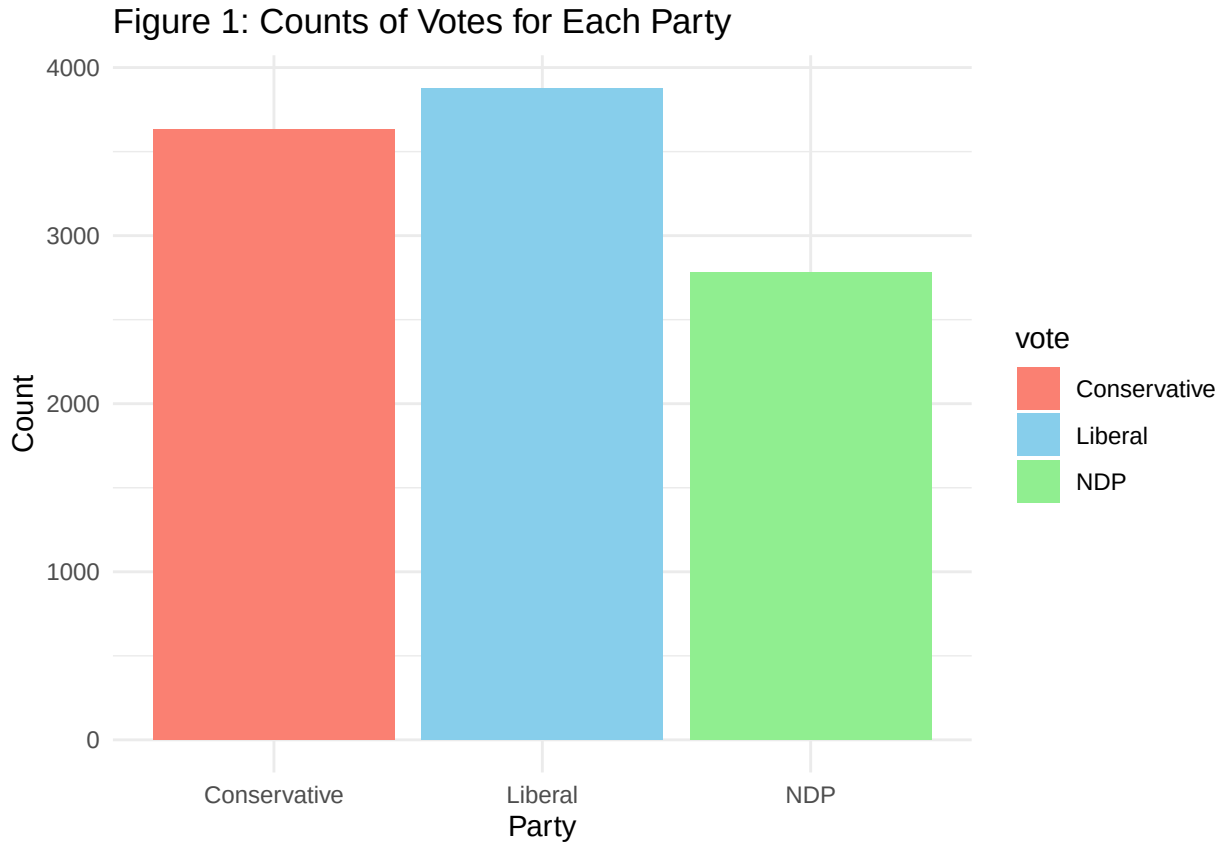
We must then tackle the ethical problem of non-representation. Although it is true that by removing these categories we risk not representing non-binary attitudes, if we tried to randomly assign non-binary responses to male and female we would be misgendering which is a worse ethical outcome. Although non-binary choices are not represented in our analysis, this doesn't prevent them from expressing themselves politically by participating in elections, it only removes their influence on voting patterns in our statistical analysis. Thus, we believe the ethical concern is not pressing to the degree that it should invalidate our decision to remove the data.

Table 1: Combined Age Data

avg_age_census	avg_age_survey
52.19011	51.39517

Table 2: Vote Summary Data

vote	n
Conservative	3631
Liberal	3879
NDP	2781



As we can see, from table 2 Vote Summary Data, there are more people who have voted for the Liberal party in the survey than the Conservative party in the survey. The liberal party has 3879 votes whereas the conservatives have around 3631. The NDP stands last at around 2781. This supports our initial hypothesis that the Liberal party has a higher vote share.

Figure 2: Proportions by Province and Sex

According to figure 2, for every province there is a difference in males and females according to survey and census statistics. For example, in Alberta there are more females in the survey data relative to census whereas for Males there are more people in census relative to survey. These differences highlight the fact that our survey may be non-representative, thus validating the need for post-stratification to weight the survey data according to the census data estimates in order to get an accurate population estimate.

Figure 3: Combined Boxplots of Age for Survey and Census Data

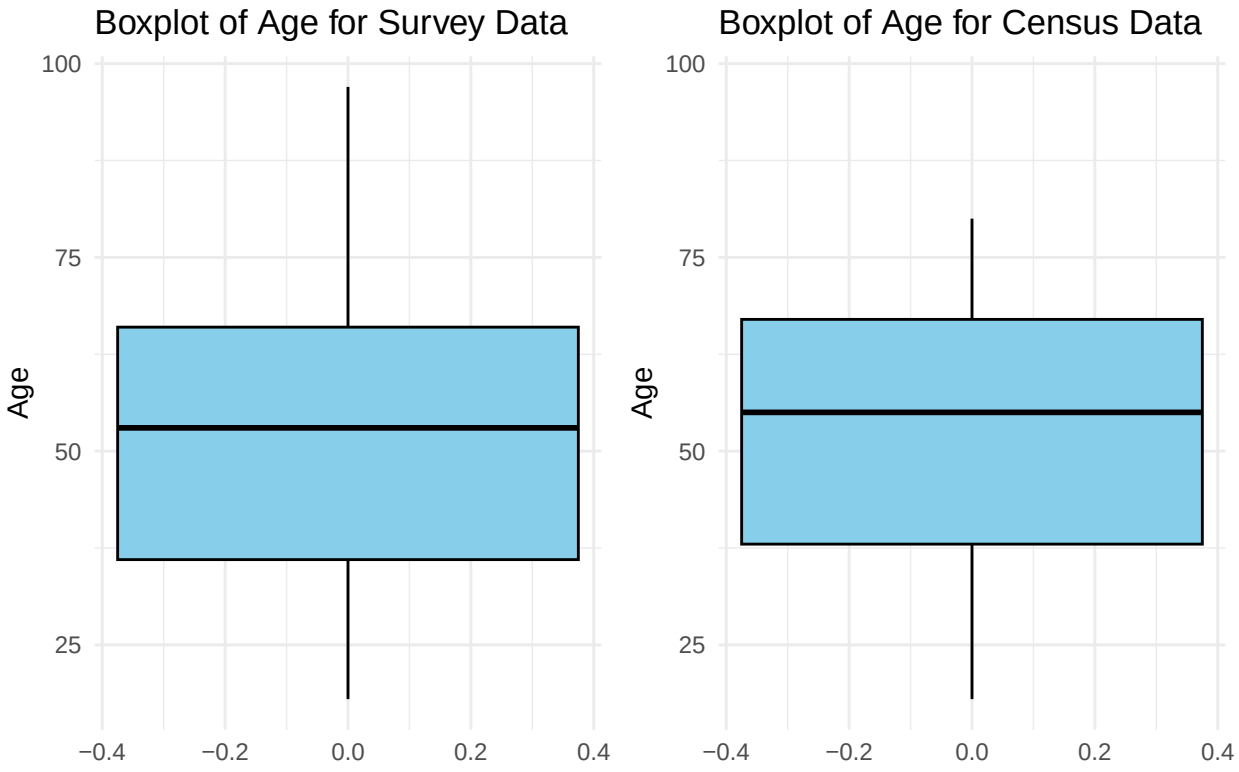


figure 3

according to the boxplot, the median age in census data is higher than that of survey data. They both seem to have similar spread suggesting that age is already reasonably well correlated with census data. It also shows that 50% of Canadians in the survey are between around 30 and 60 (with similar bounds for census data). This could affect our prediction since it implies that young voters are not as well-considered by the survey or census (ages 18-30).

Methods

Our objective is to understand the popular vote in the Next Federal election in

Canada. This is done to tackle the problem of lack of data on political preferences in the Canadian population based off demographic variables. In order to predict the popular vote, we will first understand the influence that certain voter characteristics have on the probability that someone votes for a party. For our model, we use Age,Sex,Province,language, marital status, child status, and education of voting citizens as predictors to understand how they contribute to the voting probability for a particular party. This calls for a logistic regression model where the categorical response variable is whether the survey respondent is voting or not voting for a specific party, where Yes is denoted by 1 and No is denoted by 0. We will build three logistic regression models for the dominant 3 parties: The NDP (New Democratic Party) the Conservative Party and the Liberal Party. Here is a short description of the predictor variables

-Sex: A binary categorical variable that tells us whether the respondent is either 'Male' or 'Female'.

-Age: a discrete numerical variable that tells us the age of the survey respondent.

-Province: a nominal categorical variable that tells us the province of residence for each survey respondent

-Education: A nominal categorical variable that includes the level of education which includes less than high school, high school, Bachelors, trade certificate etc.

-Language: Language spoken by respondent, either French,English or French and English.

-Marital status: Includes marital status like married, single, widowed,divorced,living as per common law etc.

-child status: Do the respondents have children yes or no (a more detailed breakdown of the model variables will be given later)

These variables were picked from the census and survey data and matched because there exists evidence to support their influence on voting preference. EKOS politics is a website with election-related data and political analysis done by Canadian polling firm EKOS Research[6] reports : Gender/sex influences voting preferences with, Women generally voting for the liberal party more while Men tend to vote for the conservative party more. Age influences voting preferences where people in age groups 50+ tend to vote for Liberal wherein the Conservative party doesn't see much variation in voting [6]. People with University degrees tend to vote more for the Liberal party whereas people with college degrees or high school education vote more for the Conservative party relative to university students. Province wise, provinces like Alberta and Saskatchewan vote highly in favour of the conservatives whereas in provinces like Ontario and Quebec the Liberals are more popular.[6]EKOS Politics. (2020)

We use a Logistic regression model because the model gives us the probability of any one event occurring from our binary response variable (in this case voting or not voting for a party). These probability figures will help us compare which party is more likely to win i.e get the popular vote based on our survey data. We will also use post-stratification which is a method of adjusting sampling weights for our sample variables of interest (age,sex,province) using population level data on age, sex ,and province to calculate our measure of interest(probability of voting) which helps improve the accuracy of our estimation.

Managing missing data when constructing the model we deleted all the rows containing missing data. The implications of this move on the analysis is as follows. Firstly, by removing missing data we are introducing bias into our model with respect to all the variables that consist of missing data, since we may be overestimating or underestimating the slope of each predictor and the intercept. Moreover, by removing the rows which have missing data, we are removing non-missing data from other columns which further hinders our ability to model an accurate relationship between our predictors and response. Thus, our analysis will be interpreted taking into consideration the variability in slopes and intercepts as a result of missing data deletion. We chose to delete the missing data because we do not have knowledge about techniques like data imputation.

Model Specifics

We have developed three logistic regression models with multiple predictors. Our binary response variable is vote_liberal, vote_conservative and vote_NDP which tells us whether the survey respondent votes for each of these parties or not. Our predictors for all three models are the same, namely age,sex, province, language, marital status, child status, and education. The Mathematical model is as follows.

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{\text{age}} + \beta_2 x_{\text{SexMale}} + \beta_{3,4,5,6,7,8,9,10} x_{\text{province}} + \beta_{11,12,13} x_{\text{languageknowledge}} + \beta_{14,15,16,17,18,19} x_{\text{Maritalstatus}} + \beta_{20,21} x_{\text{children}} + \beta_{22,23,24,25,26} x_{\text{Education}}$$

$\beta_{3,4,5,6,7,8,9,10}$ for example is the beta matrix for all indicator levelsin categorical predictor province. the $x_{province}$ is the design matrix with all the predictor values. The same goes for the other categorical variables.

i = 1,2,3 where

i = 1 represents p1 which the probability of voting liberal

i = 2 represents p2 which is the probaility of voting Conservative

i = 3 represents p3 which is the probability of voting NDP

model justification

In order to determine the optimal model, we will use model selection criterion to pick the most significant predictors from our current model. The issue at hand is that although extra predictors help explain more variation in the model, it can lead to an overfitted model which makes it less accurate. Thus, we will determine which model is an optimal balance between accuracy and number of predictors

(variation explained).

For this we will use the BIC or the Bayesian information criterion. This term takes a log likelihood of our model parameters in order to measure goodness i.e how small the residual sum of squares is. It also imposes a penalty for adding extra predictors that do not significantly reduce residual sum of squares, thus favouring simpler models. We will use a procedure called automatic backward selection. Since we are working with a complex model with 7 predictors, automation will make the task of picking good predictors easier. Backward selection is the process of taking a model with a set number of predictors and deleting them one by one on the basis of reducing overall model BIC until a model is reached with the lowest possible BIC.

Based on the backward selection, variables removed are sex,children and Marital status. This simplifies our model to include the predictors language_knowledge, education, province, and age. The removal of the persons marital status and number of children should not be a problem since our objective is to understand influential demographic variables that affect the popular vote. However, eventhough sex is removed as a variable we believe it to be an important variable of interest when it comes to understanding political preferences. This is further reinforced by the Ekos Politics report which shows differences in voting preference based on sex. Thus, the new mathematical model is the following:

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{SexMale} + \beta_{3,4,5,6,7,8,9,10} x_{province} + \beta_{11,12,13} x_{languageknowledge} + \beta_{14,15,16,17,18} x_{Education}$$

Describing parameters

-β_0 is the intercept of the model which indicates the log odds of voting for a party given that all other predictor variables are zero.

Table 3: Variance Inflation Factor (VIF) for ModelX

	vif.GVIF	vif.Df	vif.GVIF..1..2.Df..
age	1.481712	1	1.217256
province	3.060007	9	1.064105
sex	1.083500	1	1.040913
language_knowledge	3.095761	3	1.207243
marital_status	2.037824	6	1.061119
children	1.617431	2	1.127733
education	1.077994	8	1.004705

-β_1 is the change in log-odds of voting for a party per unit change in age.

-β_2 is the change in log-odds of voting for a party given that the sex is male.

-$\beta_{3,4,5,6,7,8,9,10}$ is the change in log-odds of voting for a party given each of the provinces indicated through the indicator variable $x_{province}$

-$\beta_{11,12,13}$ is the change in log-odds of voting for a party given the respondent is an english speaker, french speaker or both english and french speaker based on the indicator variable $x_{languageknowledge}$

-$\beta_{14,15,16,17,18}$ is the change in log odds of voting for a party given one of the 6 education levels based on the indicator variable -$x_{Education}$

-x_{age} is the numerical variable that indicates age

-$x_{SexMale}$ is a binary categorical variable that indicates either male or female

-$x_{province}$ is a nominal categorical variable that consists of all the provinces

-$x_{languageknowledge}$ is a nominal categorical variable that consists of the different types of language speakers.

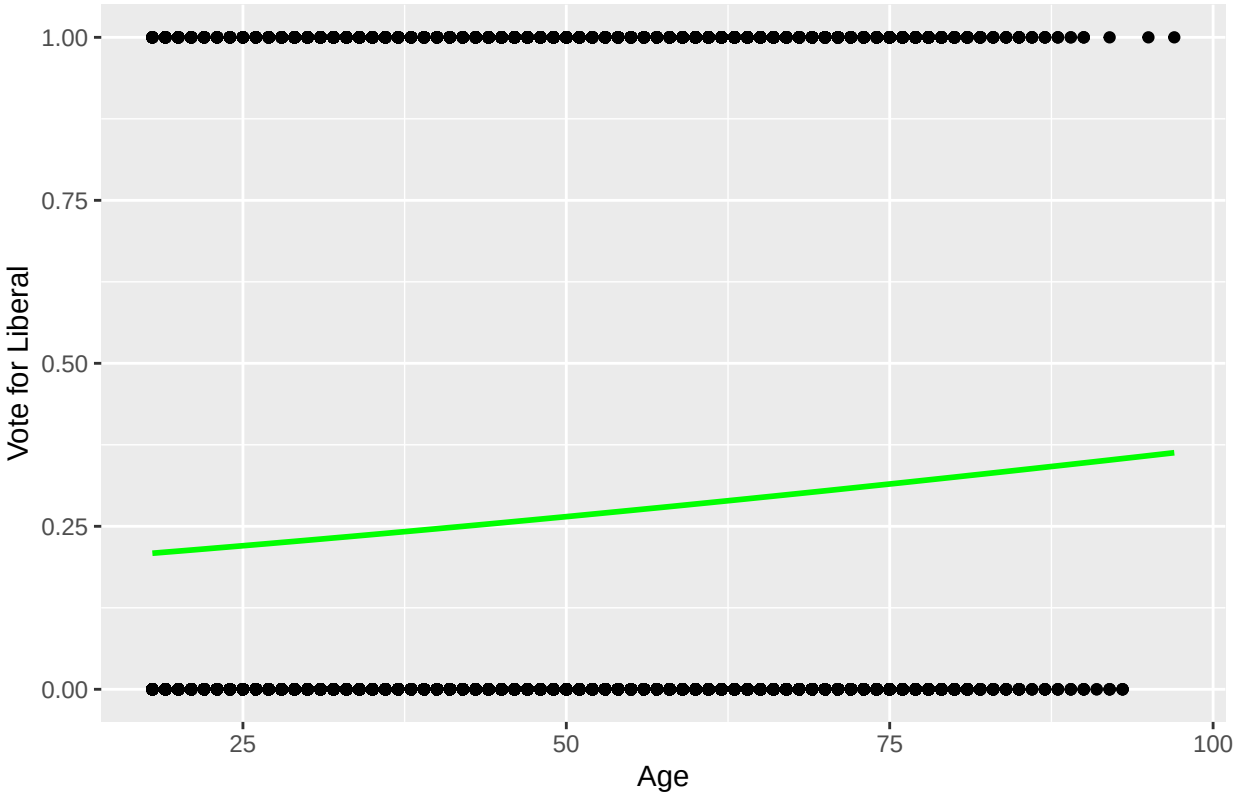
-$x_{Education}$ is a nominal categorical variable that consists of all the different education levels of the survey respondents

Testing model assumptions

For a binary logistic regression model, there are 3 main assumptions we need to look out for.Firstly, we need the response variable to be a binary categorical variable which is the case in our response variables vote_liberal, vote_conservative, and vote_NDP since the response is either 1 (vote for party) or 0 (don't vote for party). Secondly, there must be no multicollinearity between the predictor variables. For this we will use the generalized variance inflation factor to see if any large correlation exists between the predictors. Variance inflation factor is used to determine how much larger the variance of a coefficient is due to multicollinearity. If the GVIF factor is above 5 it means there is severe multicollinearity.

As we can see, none of the variables have a GVIF above 5 and so there exists no multicollinearity between the predictors. Lastly, we must check for influential outliers. This is only applicable to numerical predictors, since categorical predictors have levels as opposed to data points in a space which means there is no scope for undue influence. The following is the logistic regression for our numerical predictor age.

Figure 4: Relationship between Age and Liberal Votes



As we can see, the logistic regression curve is a smooth continuous curve. There are only 3 points that represent ages above 80 that deviate slightly from the other points but do not seem to affect the smoothness of the curve. Thus, we can say that there are no problematic outliers that may skew the results of our logistic regression. Therefore, we can say that our assumptions are satisfied.

Post-Stratification

Post-stratification is a method used when strata in our survey data is non-representative relative to population level data. strata refers to specific categories given to variables of interest. For example, sex can be said to have two strata namely male and female. Now if our variable of interest has very different strata numbers from their corresponding population strata numbers, post-stratification uses the population level strata numbers to do a weighted estimation of our measure of interest from the sample data. In simple words, we adjust our sample statistic which is the proportion of voters voting for a specific party, to the population by using census-level data. Since we are calculating probability, we will do a weighted average of each cell-level voting probability measure for cells in the population (census data). The following is the formula for the post-stratified estimator for the population proportion of people who vote for a specific party.

$\overline{p}_{PS} = \frac{\Sigma N_h \times \overline{p}_h}{\Sigma N}$

This will be repeated three times for the three parties of focus: Liberal,Conservative, and NDP. On key assumption while doing this is assuming that our data is non-representative of the population. Thus, post-stratification helps improve the accuracy of estimation of the proportion of voters that will vote for a specific party by taking population-level weights of our sample variables of interest. when post-stratifying we are considering 4 variables. Education, province, language, age, and sex. The following are the levels for each variable that will be used to construct the weighted average during post-stratification.

Education: - College, CEP, or other non-university certificates or diplomas

- High school diploma or a high school equivalency certificate
- Bachelor's degree (e.g. B.A., B.Sc., LL.B.)
- Less than high school diploma or its equivalent
- University certificate or diploma below the bachelor's level
- Trade certificate or diploma

Province

- Ontario
- Quebec
- British Columbia
- Saskatchewan
- Nova Scotia
- New Brunswick
- Prince Edward Island
- Manitoba
- Newfoundland and Labrador
- Alberta

Language - English only

- French only
- Both English and French

Age has several strata that is described in the data section.

Sex

- male - female

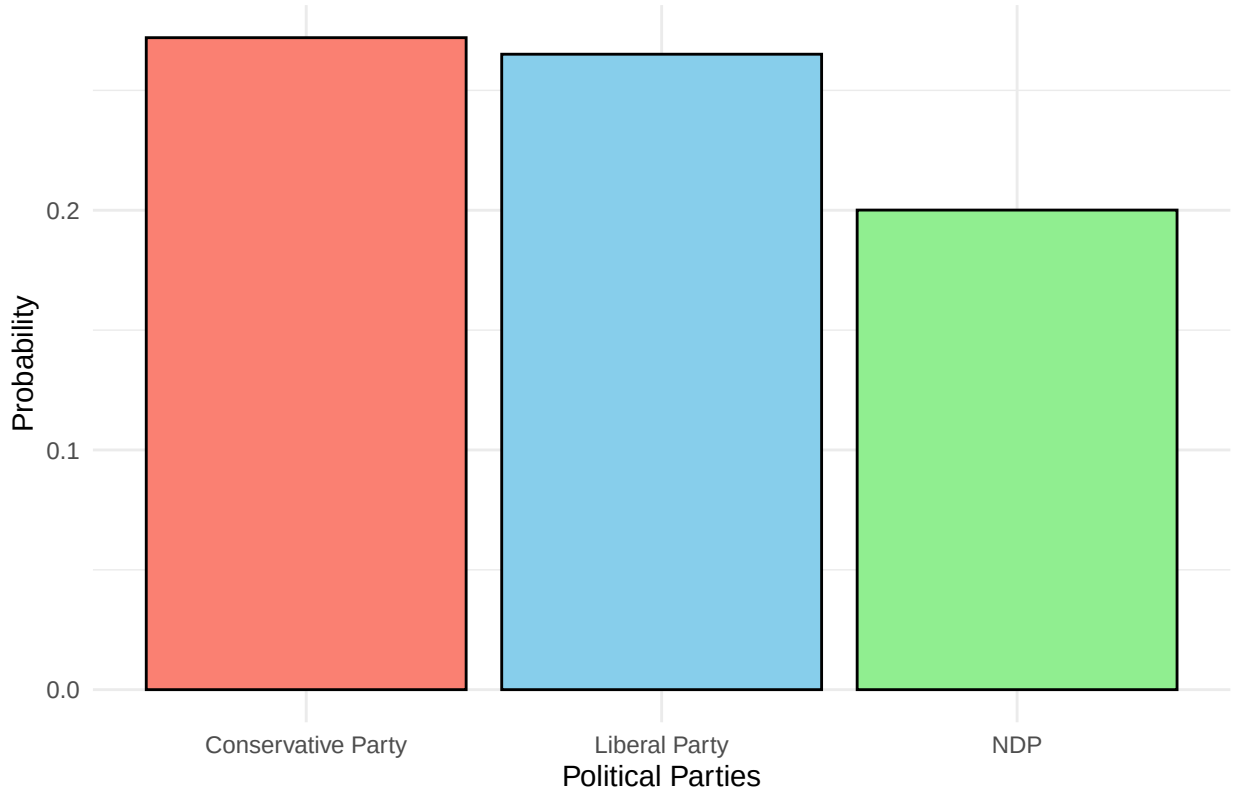
Model	Probability
Liberal Party	0.2650799
Conservative Party	0.2719771
NDP	0.2000656

Results

Analyzing the results of the poststratification, the conservative party and the liberal party are predicted to be in close running. The conservative party ended up just barely in the lead with 27.2 % , followed closely by the liberal party 26.51 %. The NDP ended up in third place with 20.01 %. The conservatives and the liberals are slated to have a difference of less than 1% in the end which shows how volatile the political climate is predicted to be come 2025. The NDP are not too far behind get 7.19% less votes than the conservatives and 6.50% behind the liberal. These results seem reasonable as most of our coefficients are highly significant and would suggest that our popular vote percentages should be reliable. Although, with the unexpected turn of the recent electoral outcomes the tides may turn again before the elections arrive in 2025. Overall, the Conservative party is predicted to secure a higher share of the popular vote by a very slight margin.

From the logistic regression, we see that Newfoundland and Labrador are highly supportive of the liberal party. The Conservative party seems unpopular in provinces like New Brunswick and Quebec. The people of Prince Edward Island are highly unlikely to vote for the NDP. Overall, age is the most significant factor deciding which party someone will vote for, with its p-value approximately 0 for all three parties. The Liberal and the Conservative party have certain province that are extremely significant on whether they'll vote for the party, like New Brunswick, Newfoundland and Labrador, Nova Scotia and Ontario for the liberals, and British Columbia, New Brunswick, Nova Scotia and Ontario for the conservatives. The most significant determinant for the NDP is whether they are male.

Figure 5: Predicted Voting Probabilities for Canadian Political Parties



Variable	Estimate	Std. Error	Statistic	Table 6: Model Summary P Value
(Intercept)	-1.6795024	0.1025528	-16.3769468	0.0000000
age	0.0120206	0.0011843	10.1496307	0.0000000
provinceBritish Columbia	0.3249940	0.0841380	3.8626297	0.0001122
provinceManitoba	0.2455509	0.1172428	2.0943803	0.0362261
provinceNew Brunswick	0.8120668	0.1452089	5.5924053	0.0000000
provinceNewfoundland and Labrador	1.0324978	0.1870565	5.5197104	0.0000000
provinceNova Scotia	0.7937342	0.1276931	6.2159538	0.0000000
provinceOntario	0.6372982	0.0678454	9.3933947	0.0000000
provincePrince Edward Island	0.3071106	0.3706013	0.8286820	0.4072844
provinceQuebec	0.4134439	0.0936953	4.4126445	0.0000102
provinceSaskatchewan	-0.4460188	0.1810506	-2.4635032	0.0137587
sexMale	-0.0287837	0.0394731	-0.7291965	0.4658815
language_knowledgeEnglish only	-0.1319292	0.0664695	-1.9848066	0.0471660
language_knowledgeFrench only	-0.4216872	0.0810794	-5.2009168	0.0000002
language_knowledgeNeither English nor French	0.1380066	0.0926226	1.4899883	0.1362273
educationCollege, CEGEP or other non-university certificate or di...	-0.4673534	0.0551927	-8.4676694	0.0000000
educationHigh school diploma or a high school equivalency certificate	-0.4327433	0.0660050	-6.5562203	0.0000000
educationLess than high school diploma or its equivalent	-0.5679940	0.1394774	-4.0723020	0.0000466
educationNA	-1.5484793	0.7475034	-2.0715347	0.0383089
educationTrade certificate or diploma	-0.4562345	0.0767957	-5.9408845	0.0000000
educationUniversity certificate or diploma below the bachelor's level	-0.2031421	0.0682097	-2.9781982	0.0028995
educationUniversity certificate, diploma or degree above the bach...	0.1039665	0.0926868	1.1216978	0.2619909
educationUniversity certificate, diploma or degree above the bachelor's level	0.0701409	0.0663203	1.0576081	0.2902341

Variable	Table 7: Model Summary Estimate	Std. Error	Statistic	P Value
(Intercept)	-1.5694168	0.1037949	-15.1203683	0.0000000
age	0.0133930	0.0012343	10.8510325	0.0000000
provinceBritish Columbia	-0.7953281	0.0784563	-10.1372131	0.0000000
provinceManitoba	-0.4865683	0.1069587	-4.5491245	0.0000054
provinceNew Brunswick	-1.1519543	0.1720232	-6.6965039	0.0000000
provinceNewfoundland and Labrador	-1.0286509	0.2207696	-4.6593866	0.0000032
provinceNova Scotia	-1.0476487	0.1435856	-7.2963359	0.0000000
provinceOntario	-0.5876119	0.0596048	-9.8584699	0.0000000
provincePrince Edward Island	-0.5519935	0.3436723	-1.6061625	0.1082382
provinceQuebec	-1.1122958	0.0957211	-11.6201739	0.0000000
provinceSaskatchewan	0.0906432	0.1283982	0.7059541	0.4802167
sexMale	0.5048719	0.0408447	12.3607637	0.0000000
language_knowledgeEnglish only	0.1790034	0.0731471	2.4471709	0.0143983
language_knowledgeFrench only	-0.1380404	0.0941674	-1.4659035	0.1426746
language_knowledgeNeither English nor French	0.3152212	0.1007120	3.1299271	0.0017485
educationCollege, CEGEP or other non-university certificate or di...	0.2656640	0.0558504	4.7567098	0.0000020
educationHigh school diploma or a high school equivalency certificate	0.2738931	0.0660151	4.1489498	0.0000334
educationLess than high school diploma or its equivalent	0.3330833	0.1285473	2.5911343	0.0095660
educationNA	-0.0415123	0.5610105	-0.0739956	0.9410139
educationTrade certificate or diploma	0.2003382	0.0763603	2.6235926	0.0087008
educationUniversity certificate or diploma below the bachelor's level	0.0118349	0.0739117	0.1601216	0.8727853
educationUniversity certificate, diploma or degree above the bach...	-0.1506596	0.1066940	-1.4120723	0.1579287
educationUniversity certificate, diploma or degree above the bachelor's level	-0.1905477	0.0769748	-2.4754546	0.0133067

Table 8: Model Summary				
Variable	Estimate	Std. Error	Statistic	P Value
(Intercept)	0.4606606	0.1075097	4.2848292	0.0000183
age	-0.0321726	0.0013957	-23.0506588	0.0000000
provinceBritish Columbia	0.4385022	0.0845398	5.1869299	0.0000002
provinceManitoba	0.2026082	0.1205455	1.6807607	0.0928094
provinceNew Brunswick	-0.0190177	0.1734072	-0.1096709	0.9126704
provinceNewfoundland and Labrador	0.1324844	0.2147967	0.6167898	0.5373734
provinceNova Scotia	0.2821637	0.1385451	2.0366194	0.0416882
provinceOntario	-0.0784924	0.0709417	-1.1064350	0.2685383
provincePrince Edward Island	-1.7144086	0.7354215	-2.3311920	0.0197432
provinceQuebec	-0.4871567	0.1083964	-4.4942146	0.0000070
provinceSaskatchewan	0.2623203	0.1464413	1.7913004	0.0732451
sexMale	-0.3714967	0.0469249	-7.9168441	0.0000000
language_knowledgeEnglish only	-0.0790366	0.0724668	-1.0906593	0.2754228
language_knowledgeFrench only	-0.4969977	0.0981419	-5.0640710	0.0000004
language_knowledgeNeither English nor French	-0.6297994	0.1193359	-5.2775367	0.0000001
educationCollege, CEGEP or other non-university certificate or di...	0.0540114	0.0627345	0.8609521	0.3892644
educationHigh school diploma or a high school equivalency certificate	-0.0591741	0.0783953	-0.7548175	0.4503585
educationLess than high school diploma or its equivalent	-0.0032343	0.1607766	-0.0201169	0.9839501
educationNA	-0.1885322	0.5705138	-0.3304604	0.7410521
educationTrade certificate or diploma	0.2387438	0.0834978	2.8592821	0.0042460
educationUniversity certificate or diploma below the bachelor's level	0.1388143	0.0764067	1.8167801	0.0692508
educationUniversity certificate, diploma or degree above the bach...	-0.1015462	0.1234704	-0.8224341	0.4108299
educationUniversity certificate, diploma or degree above the bachelor's level	0.1862329	0.0797382	2.3355549	0.0195145

Conclusions

Summary of the Hypotheses, Methods and Results

We started with the hypothesis that the Liberal party would secure the popular vote. This hypothesis was made on the basis of polling data from 2020 by EKOS politics which indicated that the Liberal party largely received the popular vote across several categories like age, gender, province etc. In order to assess this hypothesis, we sought to use data from the Canadian Election Study 2021 by the Consortium on Electoral Democracy to determine the popular vote in the next federal elections. In order to do this, we chose to use a multiple logistic regression which consisted of 7 predictors initially. These predictors were selected on the basis of their observed affect on voting preference in the 2020 polling study by EKOS politics. Three logistic regression models were constructed with three binary response variables vote liberal, vote conservative ,and vote NDP. We then used the automated backward selection and the BIC criterion to find an optimal set of predictors that maximizes the goodness of the model (minimizes residual sum of squares) while incentivizing a simple model with less predictors over a complex one. Based on this method, we landed up with a 5 predictor logistic regression model with Age,Sex,Language, Province,and Education. Then, we conducted a post-stratification on all three models using census-level data where we did a weighted average of voting probability for each cell weighted by the population level strata for each predictor variable provided in the census data. By doing this, we got the post-stratified probability of people voting for the Liberals, Conservatives, and NDP. The results we got is that (insert figure) is the probability that people would vote for the Conservative party. (insert figure) is the probability that people will vote the Liberal party and (insert figure) is for the NDP.

Key results Overall, the Conservative party has the highest share of the popular vote as per the survey data and our model assumptions although by only a small margin. This successfully answers our initial research question of finding the overall popular vote for the 2025 Federal election in Canada. We selected the Liberal, Conservative and NDP only as parties because they have repeatedly dominated the Canadian political space.

Analysis and big picture Two aspects are interesting about the result. One, the Conservative party now seems to have the highest share of the popular vote and two, the differences in the share of the popular vote are very small.

The gaining momentum of the conservative party can be explained through rising 'economic anxiety' where people are feeling the financial pressure from a cost of living crisis. According to polls by Nanos research 'Jobs, inflation and the cost of housing' has been the top 5 priority issues among Canadians that the Conservative party has been talking about. People likely associate these problems with the policies of the Liberal party

given that the Conservatives have a lead[2].

The fact that the differences between Conservatives and Liberals are small indicates that there is still significant support for the Liberal party. the NDP are around 7% down relative to the Liberals and Conservatives as per our model which is similar to the situation in the 2021 elections where the Liberals had to form a coalition with the NDP in order to form a majority. According to this article from the Global News, polls show that the Liberal party is at 32% of votes whereas the conservative are at 37% whereas the NDP is at 16%. This shows that our results underestimate the conservative lead as per current polls while it overestimates the NDP voteshare in relation to current polls. Thus, we can say that the NDP has fallen in popularity and that the Conservatives have surged forward[3].[2]Spencer Van Dyk. (2023, August 31). Conservatives maintain months-long lead over liberals, as Canadians' economic anxiety rises: Nanos. CTVNews. [3]Aziz, S. (2023, July 4). Conservatives edge ahead of liberals in voter support, breaking gridlock: Poll. Global News. https://globalnews.ca/news/9808274/conservative-liberal-voting-intention-ipsos-poll/

Weaknesses One limitation of our study is the fact that Census data does not include Gender. This essentially prevents us from understanding the voting patterns of non-binary and transgender peoples which treads on the issue of non-representation.

Another issue is that missing data has been deleted which has undoubtedly removed observations that may otherwise be important for our analysis. Thus, our values for voting probability will fall within an interval of error.As observed above when comparing our results with Global News's reporting on current polls, the differences seen in voteshare for each party between our model and the polls could be in part due to missing data. Additionally, in order to match the survey data to the census data certain adjustments were done like clubbing in 'some college CEGP or other..finished' in the survey data to trade certificates in the census data. Even in Language data, we removed people who responded 'did not know'(only 7 observations so would not have a big impact). All this naturally increases the interval of error for our estimates both in our logistic regression and post-stratification.

Next Steps A good continuation of this project would be to use more sophisticated models like multinomial models and multilevel logistic regression models to factor in mixed effects.This could help us better model population level trends using inference. Furthermore, imputation techniques need to devised and used to solve the sex to gender problem and the missing value problem. In conclusion, we have successfully produced overall vote-share estimates that reasonably model current polling data with some variation due to differences in time between the survey data and polling data along with methodological reasons relating to missing data, Sex and Gender and model assumptions. This analysis can form a benchmark for future analysis on the political preferences of Canadian citizens using more recent polling data and more advanced regression methodologies.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: April 4, 1991)

2. RStudio Team. (2020). *RStudio: Integrated Development for R*. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

3. Allaire, J.J., et. el. *References: Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: April 4, 1991)

4. OpenAI. (2023). *ChatGPT (September 13 version) [Large language model]*. https://chat.openai.com/chat (Last Accessed: September 13, 2023)

5. Aziz, S. (2023, July 4). Conservatives edge ahead of liberals in voter support, breaking gridlock: Poll. Global News.

https://globalnews.ca/news/9808274/conservative-liberal-voting-intention-ipsos-poll/

6. EKOS Politics. (2020). UPDATE ON THE POLITICAL LANDSCAPE AND THE ISSUES OF RACE, POLICING, AND THE THREE MS IN THE CANADACHINA AFFAIR. https://www.ekospolitics.com/wp-content/uploads/full_report_june_26_2020a.pdf

7 . Spencer Van Dyk. (2023, August 31). Conservatives maintain months-long lead over liberals, as Canadians' economic anxiety rises: Nanos. CTVNews. https://www.ctvnews.ca/politics/conservatives-maintain-months-long-lead-over-liberals-as-canadians-economic-anxiety-rises-nanos-1.6541736

8. All analysis for this report was programmed using `R version 4.0.2` [2].

Appendix

Generative AI Statement

This response was generated using ChatGPT by OpenAI, (ChatGPT 2023) Version 3.5. Accessed November 23rd, 2023.

- I used the prompt "explain why age, province, sex, language, marital, children and education would be indicators of what party someone would vote for." The prompt significantly aided us in structuring our rationale behind utilizing the predictors, contributing to clearer articulation, refining grammar, and ensuring accuracy in spellings.
- I used the prompt "how to interpret GVIF and what is the threshold of multicollinearity for GVIF."This prompt was instrumental in clarifying the interpretation of GVIF, the indicator for multicollinearity in regression models, and helped define its critical threshold value.
- I used the prompt "how to take a specific obervation from a table and assign it to x" This prompt was essential in acquiring a precise value, aiding in the integration of necessary inline code within the report,
- I used the promt " explain Error: unexpected '/' in "options(repos = structure(c(CRAN = https:/"" facing this error this is my code " options(repos = structure(c(CRAN = https://cran.r-project.org/))) install.packages("cowplot") install.packages("kableExtra")" This code was pivotal in resolving the underlying issue behind the code malfunction in plotting Figure 3, providing clarity on the primary cause of the problem.
- I used the promt "exaplin this Missing inserted. l.480 log (frac{p_i}{1-p_i}right) = textbackslash{} Try to find the following text in Assignment2-starter_code.Rmd: logleft You may need to add around a certain inline R expression r in Assignment2-starter_code.Rmd (see the above hint). See https://github.com/rstudio/rmarkdown/issues/385 for more info. Error: LaTeX failed to compile Assignment2-starter_code.tex. See https://yihui.org/tinytex/r/#debugging for debugging tips. See Assignment2-starter_code.log for more info. Execution halted explain this error" Once more, this code was instrumental in deciphering the error encountered while configuring LaTeX for the mathematical model, aiding in a better understanding of the issue faced.
- I used the prompt " explain this error to me summarise()`has grouped output by 'province'. You can override using the`.groups`argument.`summarise()`has grouped output by 'province'. You can override using the`.groups' argument. Encountering this error within Figure 2 of the grouped graphs prompted quick resolution upon its identification, resulting in a swift solution.