Forecasting Canada's 2025 Federal Election: A Regression Analysis with Post-Stratification Approach

STA304 - Fall 2023 -Assignment 2

Labib Zaman

November 22, 2023

Introduction

Predicting election results is a crucial component of political science and data analytics, influencing not just academic discussion but also public policy and democratic participation. This research looks into a prediction analysis of the overall popular vote for the approaching 2025 Canadian federal election. The study attempts to predict election results by using regression models with post-stratification, utilizing a large dataset from the General Social Survey (GSS) and the Canadian Election Study (CES).

Such an examination is important for reasons beyond scholarly interest. Through the emphasis of voter preferences and societal trends, it provides insights into the changing political scene of Canada. This becomes more important when considering Canada's dynamic and varied electorate, where socioeconomic, demographic, and regional variables interact to influence political views.

To provide a foundation for understanding, here are some political terminologies that will be used throughout the report:
Electorate - all the people in a country eligible to vote in an election

The central research question guiding this study is: "Can the overall popular vote in the 2025 Canadian federal election be accurately predicted using a regression model with post-stratification based on current socio-demographic data?"

Based on the GSS and CES datasets, we hypothesize that a combination of demographic, socioeconomic, and geographic factors will be able to forecast the winning party in the Canadian federal election of 2025 in terms of the popular vote. We predict that a number of factors, including age, education, region, and past voting behaviour, will be important predictors of the majority party choice among Canadian voters.

Data

The General Social Survey (GSS) and the Canadian Election Study (CES) are two important sources of data used in this study. The Leger Opinion panel was used to gather a representative online sample of 20,968 Canadian citizens for the CES data. The sampling was carefully planned to include a wide range of demographics from Atlantic Canada to the Western provinces and the Territories. It was stratified by area, balanced for gender and age within each region, and representative. Simultaneously, similar to census data, the GSS data provides a more comprehensive and general demographic picture of the Canadian population, which is useful for reweighting and altering the CES data in order to better reflect the population as a whole.

No additional cleaning was done for the GSS data beyond the initially cleaned data.

Cleaning Process: Combining and Aligning Datasets: To guarantee consistency, three important characteristics have been unified across both datasets: age, province, and education level. This procedure comprised:

1. Variable Alignment: Recoding the education categories in both datasets to create a common set of categories. More specifically, for the following answers:

- No schooling (1) – 'Less than High School'
- Some elementary school (2) – 'Less than High School'
- Completed elementary school (3) – 'Less than High School'
- Some secondary/ high school (4) – 'Less than High School'
- Completed secondary/ high school (5) – 'High School Diploma or Equivalent'
- Some technical, community college, CEGEP, College Classique (6) – 'Some College/University'
- Completed technical, community college, CEGEP, College Classique (7) – 'Completed College/University Below Bachelor's Level'
- Some university (8) – 'Some College/University'
- Bachelor's degree (9) – 'Bachelor's Degree'
- Master's degree (10) – 'Above Bachelor's Degree'
- Professional degree or doctorate (11) – 'Above Bachelor's Degree'
- Don't know/ Prefer not to answer (12) – Omitted from the survey

Similarly for the GSS Data, mostly renaming for simplicity purposes:

- Less than high school diploma or its equivalent – 'Less than High School',
- High school diploma or a high school equivalency certificate – 'High School Diploma or Equivalent'
- College, CEGEP or other non-university certificate or diploma – 'Completed College/University Below Bachelor's Level'
- Trade certificate or diploma – 'Completed College/University Below Bachelor's Level'
- University certificate or diploma below the bachelor's level – 'Completed College/University Below Bachelor's Level'
- Bachelor's degree (e.g. B.A., B.Sc., LL.B.) – 'Bachelor's Degree',
- University certificate, diploma or degree above the bachelor's level – 'Above Bachelor's Degree'
- NA and Blanks were omitted from the survey

2. Rounding Age: In the GSS dataset, ages were rounded to match the age format in the CES dataset.
3. Province Consistency: Ensuring the province names/categories were consistent across both datasets.

Important Variables:

- **Age**: This variable is essential to understanding trends and distributions in the population. Age has a significant role in election projections since different age groups frequently display different voting behaviours.
- **Province**: The geographic variable that reflects Canada's regional differences. Voting preferences may be greatly influenced by provincial allegiance, which reflects regional political environments and concerns.
- **Education Level**: Political views and awareness are frequently associated with education. This variable aids in the comprehension of the relationship between voting behaviour and education levels.
- **Vote Choice**: The main result of interest for forecasting the popular vote is the respondents' party preference, which is captured by this variable. It includes category information that shows which political party each respondent preferred or leant toward. Predicting the election's general vote trends requires an understanding of the distribution and the variables driving these decisions.

Table 1: Age Distribution of Respondents

Mean Age	Standard Deviation
50.7652	17.0842

The mean age of the respondents in our dataset is approximately 50.77 years, with a standard deviation of 17.08 years. This suggests a broad age distribution among the participants, indicating a wide range of adult age groups represented in the survey. The significant standard deviation points to a diverse sample in terms of age.

Table 2: Distribution of Respondents by Province

Province	Number of Respondents
Ontario	5183
Quebec	4366
Alberta	1728
British Columbia	1574
Manitoba	551
Nova Scotia	359
Saskatchewan	309
New Brunswick	272
Newfoundland and Labrador	136
Prince Edward Island	42
Yukon	20
Northwest Territories	9
Nunavut	4

This table shows the distribution of respondents across Canadian provinces. The largest group comes from Ontario, representing 5183 respondents, followed by Quebec with 4366 respondents. The least represented provinces are Yukon with 20 respondents and Nunavut with only 4 respondents. This distribution reflects a concentration of respondents in Canada's most populous provinces, as expected.

Table 3: Distribution of Respondents by Education Level

Education Level	Number of Respondents
Bachelor's Degree	4281
Completed College/University Below Bachelor's Level	3160
Some College/University	2823
Above Bachelor's Degree	2072
High School Diploma or Equivalent	1871
Less than High School	346

In terms of education, the majority of respondents have a Bachelor's Degree (4281 respondents), followed by those who have completed College/University Below Bachelor's Level (3160 respondents). Respondents with education levels 'Some College/University' and 'Above Bachelor's Degree' also form significant portions of the sample, with 2823 and 2072 respondents, respectively.

Table 4: Distribution of Vote Choices Among Respondents

Vote Choice	Number of Respondents
Liberal Party	3888
Conservative Party	3633
NDP	2824
Don't know/ Prefer not to answer	2082
Bloc Québécois	1294
Another party	480
Green Party	352

The vote choice distribution indicates that the Liberal Party and the Conservative Party are the most preferred choices among the respondents, with 3888 and 3633 respondents respectively favoring them. The NDP follows with 2824 respondents. The least favored options are the Green Party and ‘Another party’, with 352 and 480 respondents respectively. A significant number of respondents, 2082, chose ‘Don’t know/ Prefer not to answer’, highlighting a degree of uncertainty or indecision among the electorate.

Figure 1: Age Distribution in the Sample

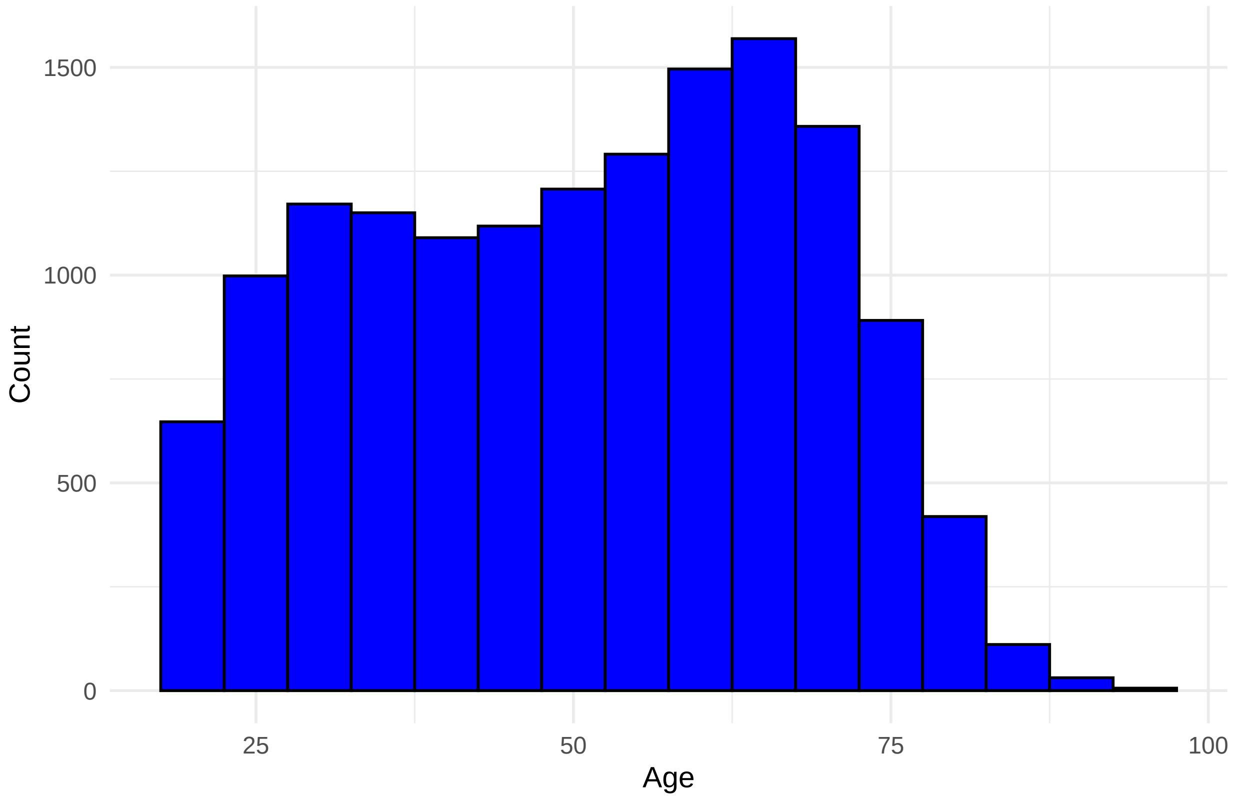


Figure 2: Age Distribution in the Population

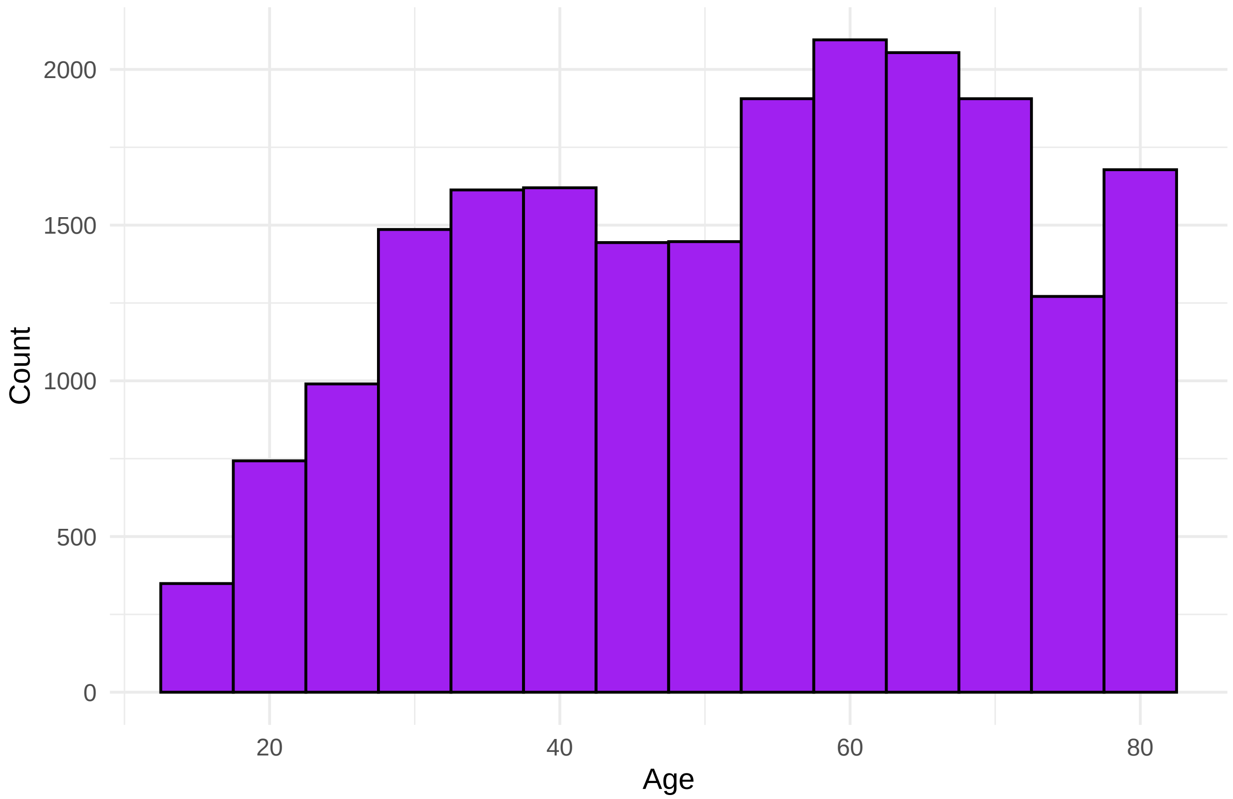


Figure 1 and Figure 2 both are histograms of age distribution but Figure 1 is in the Sample whereas Figure 2 is in the Population. Both show a relatively uniform spread across different age groups, with a slight increase in frequency as age progresses. This suggests a mature voting population, with a significant proportion of voters in the middle to older age brackets. The sample's age distribution exhibits a decline in frequency among respondents beyond the 70-75 age bracket, suggesting an underrepresentation of older age groups in the sample. In contrast, the population distribution maintains higher counts in these older age ranges, indicating that these age groups are more prevalent in the general population than they are in our sample.

There are significant ramifications for our study from this discrepancy. The influence of elderly voters on the popular vote may be underestimated due to their underrepresentation in the sample. Older age groups might differ in their political inclinations and voting habits, therefore their reduced representation in the sample may skew the results. In order to solve this, we may think about utilizing post-stratification techniques to account for the differences in the age distribution or adding weights to the sample data to better represent the age structure of the population.

Methods

Predicting the distribution of the popular vote in the next Canadian federal election is the main objective of this investigation. In order to do this, we will use a post-stratification approach in addition to a regression model to account for the representativeness of our sample with regard to the whole population. We may improve the accuracy of our prediction by modifying our projections according to demographic and geographic distributions thanks to this technique.

Model Specifics

For our regression analysis, we will use a multinomial logistic regression model, as our dependent variable voter choice is categorical with more than two levels. This model is appropriate when predicting the probabilities of the different possible outcomes of a categorically distributed dependent variable, such as the choice of political party in an election:

$$\log\left(\frac{P(Y_i = j)}{P(Y_i = k)}\right) = \beta_{0j} + \beta_{1j} x_{age_i} + \beta_{2j} x_{province_i} + \beta_{3j} x_{education_i} + \epsilon$$

Here, Y_i represents the categorical outcome of the vote choice for the i^{th} respondent, j and k index the possible outcomes (where k is the reference category), X represents the independent variables (age, province, education level), β_{0j} is the intercept term for outcome j, β_{nj} are the coefficients for each predictor variable for outcome j, and ϵ_i is the error term for the i^{th} respondent.

Any rows with missing data in the key variables were removed. This approach was chosen because post-stratification requires complete data in the key variables, and even after excluding instances that were not completed, there is still a sizable dataset that remains. Lastly, removing the missing data is assumed to be random therefore reducing the likeliness of introducing bias.

The model assumes the independence of irrelevant alternatives, that the predictors have a linear relationship with the log odds of the outcome, and that the error terms follow a logistic distribution. The parameters of interest are the β coefficients, which indicate the strength and direction of the association between each predictor variable and the log odds of voting for each party.

The multinomial logistic regression model is justified by the categorical nature of the dependent variable and the need to account for multiple predictors that vary across individuals. Post-stratification is selected due to its effectiveness in correcting for sampling bias.

Post-Stratification

Post-stratification is a statistical technique used to adjust survey results so that they better represent a known population. This procedure is essential when, as is frequently the case, the sample selected from a survey does not precisely reflect the demographic composition of the full population. Post-stratification is the process of grouping the population and the sample into uniform units, often called "cells," according to important characteristics like geographical location, age, and education.

To estimate the proportion of voters who will vote for a specific party, we first identify the distribution of our key variables within the population, typically using census data. For example, we determine the proportion of the population that falls into each age group, educational category, and province.

We then calculate the proportion of survey respondents within these same cells. If a certain cell is underrepresented in the survey compared to the population, the responses from that cell will be weighted more heavily in the final analysis. Overrepresented cells, on the other hand, will have a lower weight.

$$\hat{y}^{PS} = \sum_{c=1}^{C} w_c \cdot \hat{y}_c$$

Here, $\hat{y}^{PS}$ is the post-stratified estimate of the population mean, C is the total number of cells, w_c is the weight for cell c (calculated as the population proportion divided by the sample proportion for that cell), and $\hat{y}_c$ is the mean response within cell c in the survey.

For our analysis, we have chosen to post-stratify based on age, province, and education because these variables are likely to influence voter outcome. Age can reflect different generational perspectives, province can capture regional political climates, and education levels may correlate with political awareness and preferences. By contrast, we exclude variables that are either not available in the census data or are unlikely to influence voting behavior, such as age when first child. Including such variables would not contribute meaningfully to the accuracy of our estimates and could unnecessarily complicate the post-stratification process.

All analysis for this report was programmed using `R version 4.0.2`.

Table 5: Coefficients from the Multinomial Logistic Regression Model

	Intercept	Age	Province	Bach. Degree	Coll./Univ. Below Bach.	High School Dip.	Less than High School	Some College
2	-0.22	0.00	-0.06	0.16	0.69	0.63	0.81	0.45
3	1.79	-0.04	-0.07	-0.01	0.41	0.28	0.44	0.41
4	-15.87	0.01	1.37	0.03	0.36	0.50	0.71	0.29
5	-0.68	-0.03	-0.05	-0.22	0.15	0.01	0.21	0.24
6	-1.39	-0.02	-0.03	-0.12	0.96	0.87	1.16	0.47
7	-0.40	-0.01	0.00	0.11	0.56	0.62	0.61	0.40

Table 6: Post-Stratification Vote Distribution Estimates

Liberal Party	Conservative Party	NDP	Bloc Québécois	Green Party	Another Party	Don't Know/Prefer not to Answer
0.17	0.21	0.14	0.01	0.02	0.03	0.1

Results

In this section, we present the findings from our multinomial logistic regression analysis and post-stratification estimates. The distribution of the popular vote in the approaching Canadian federal election may be predicted in large part thanks to these results.

Regression Results

The coefficients from the multinomial logistic regression model, as shown in Table 5, elucidate the relationships between demographic factors and the likelihood of voting for each party relative to the baseline category (the Liberal Party). Notably, age and province are significant predictors of voting preferences. For instance, every additional year of age slightly decreases the likelihood of voting for the Conservative Party (by -0.06) and the NDP (by -0.07), but increases it for the Bloc Québécois (by 1.37). A key factor in determining voting behaviour is the level of education. The chance of voting for the Conservative Party rises with a bachelor's degree (coefficient of 0.16), whereas voters with less education than a high school diploma are more likely to vote for smaller parties or not at all (coefficients of 0.81 and higher).

These coefficients offer a numerical representation of the impact that every predictor has on the result. Younger voters may be more drawn to the Conservative Party and NDP, according to the inverse association shown by the negative age coefficients. On the other hand, the Bloc Québécois' positive coefficient indicates a growing chance of support among older populations.

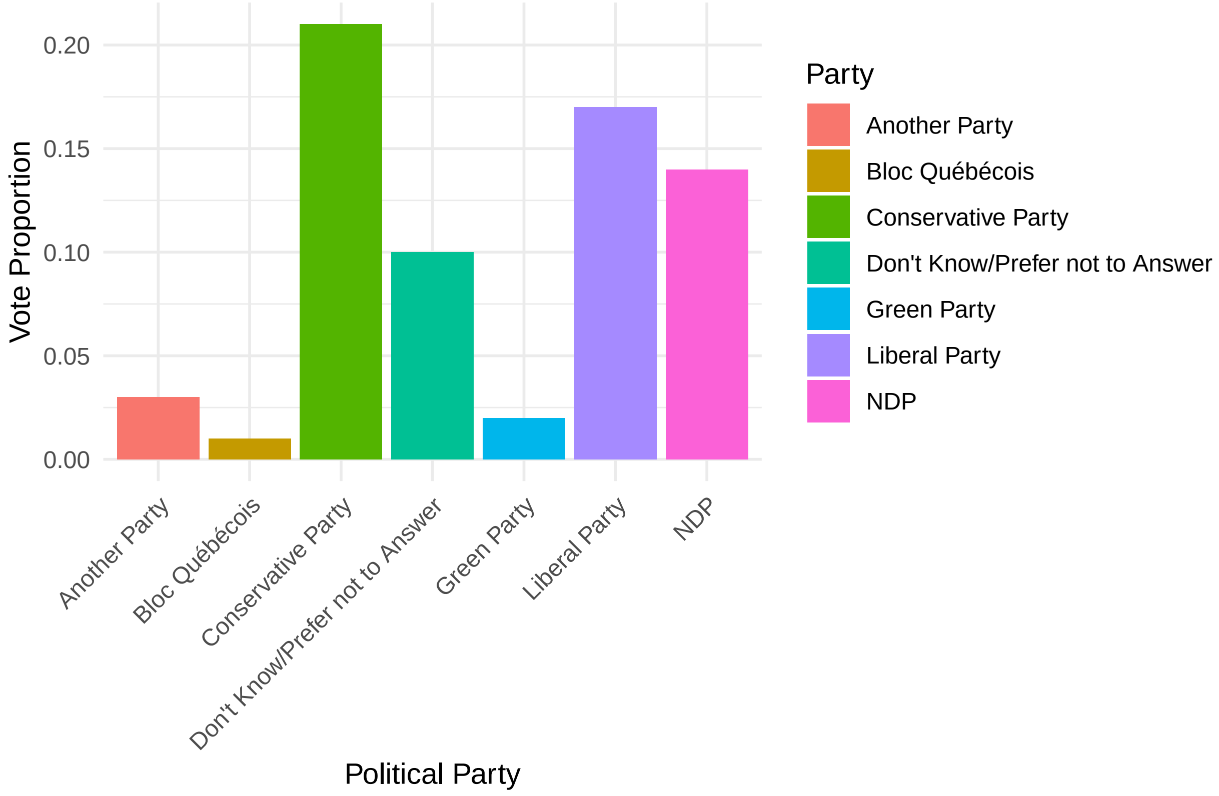
The education-related coefficients indicate a nuanced relationship between educational levels and party preference. The positive coefficients for the Conservative Party among those with a Bachelor's Degree or a college/university education below the bachelor's level suggest a correlation between higher education and preference for this party. Conversely, a high school diploma or equivalent seems to have little effect on the likelihood of voting for the Conservative Party, indicating that education may interact with other factors such as income or profession.

Post-Stratification Results

Our post-stratification analysis, presented in Table 6, estimates the proportion of votes that each major political party is likely to receive. According to our model, the Liberal Party is expected to receive approximately 17.2% of the vote, while the Conservative Party is estimated to secure about 20.9%. The NDP may receive around 13.7% of the vote. Smaller parties like the Bloc Québécois and the Green Party are estimated to receive significantly lesser proportions of the vote, at 0.6% and 1.6%, respectively. An additional 2.8% of the electorate may vote for another party, and about 10.3% are undecided or prefer not to answer.

This distribution provides insight into the current political landscape and suggests areas of strength and weakness for each party. The relatively higher estimates for the Liberal and Conservative Parties indicate a continued dominance in Canadian federal politics.

Figure 5: Predicted Vote Distribution by Party



From the chart, we can observe that the Conservative Party has the tallest bar, indicating it is predicted to receive the highest proportion of votes, roughly 20.9% as stated in your results. This is followed by the Liberal Party and the NDP, with substantial proportions as well, suggesting these parties hold significant sway over the electorate. The Bloc Québécois and the Green Party, represented by much shorter bars, are predicted to receive a smaller share of the vote.

An interesting note is the "Don't Know/Prefer not to Answer" category, which has a relatively high bar, signifying a considerable portion of the electorate who are either undecided or unwilling to disclose their preference. This highlights an area of uncertainty in the election outcome, as these votes could shift the balance if swayed towards any of the major parties. The "Another Party" category has the shortest bar, suggesting that parties outside of the main contenders are predicted to receive a small fraction of the vote.

Overall, this chart reinforces the narrative of a competitive race primarily between the Conservative and Liberal Parties, with the NDP also holding a significant presence. The data also underscores the potential impact of the undecided voters on the final election results.

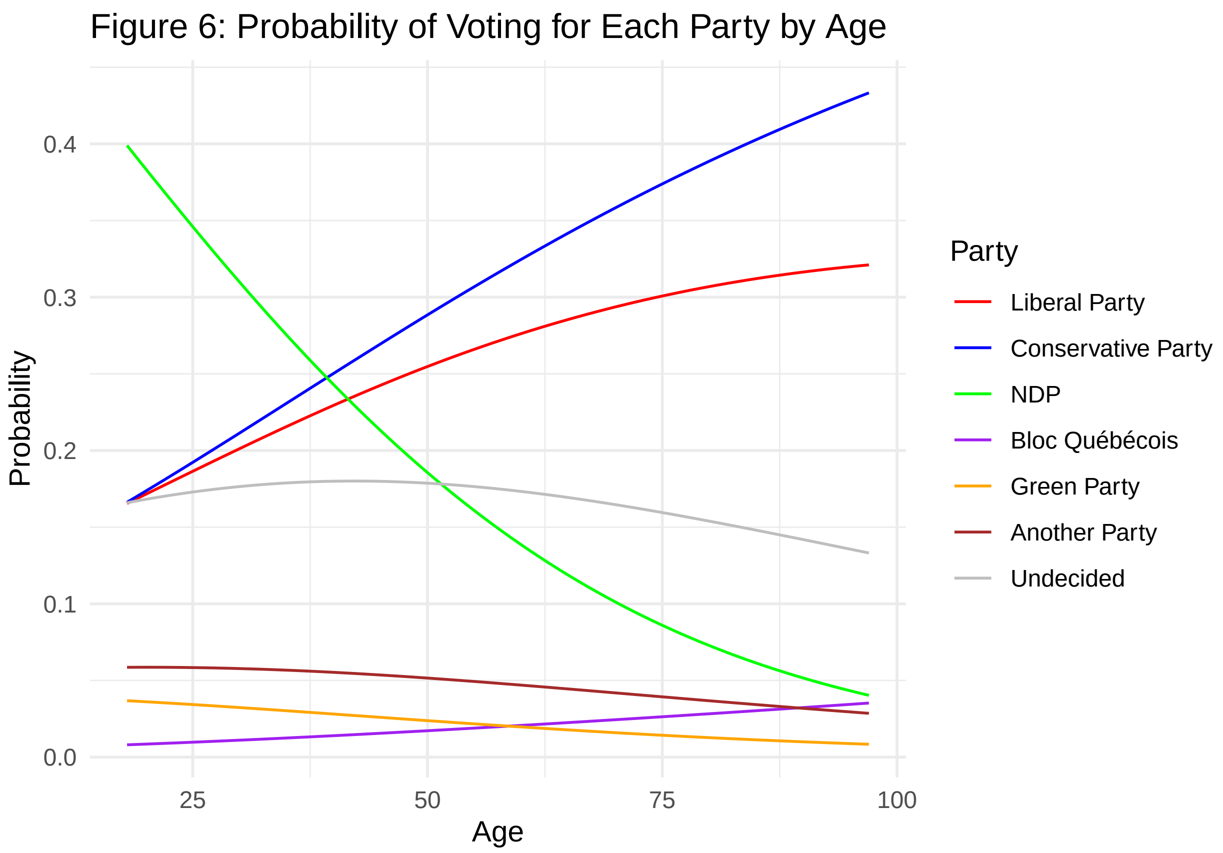


Figure 6 illustrates the probability of voting for each political party by age, showing how voter preferences change across different age groups. The likelihood of voting for the Liberal Party (red line) increases with age, suggesting that this party has more appeal among older voters. Conversely, the probability of voting for the NDP (green line) decreases as age increases, indicating that this party is more popular among younger voters. The Conservative Party (blue line) has a relatively stable probability across the age spectrum, with a slight increase in the middle age range, showing a broad appeal that does not vary drastically with age. The Bloc Québécois (purple line) has a lower overall probability compared to the major parties, with slight variations across age groups. The Green Party (orange line), Another Party (brown line), and the Undecided (grey line) have consistently low probabilities across all ages, suggesting these options are less popular or that voters are less certain about choosing them.

The graph assumes that the probability of voting for each party is a function of age alone, which may not account for other influential factors such as socioeconomic status, education, or regional issues. The steady probability for the Conservative Party across age groups may reflect the model's assumption of independence of irrelevant alternatives, meaning the choice of one alternative does not depend on the presence or absence of other alternatives. This assumption is often violated in real-world voting behavior, as the decision to vote for one party can be highly contingent on the perceived viability or policies of others. The declining probability of voting for the NDP with age could reflect generational differences in political priorities or campaign strategies that resonate more with younger voters. The model suggests that age is a significant predictor for the Liberal Party, which could be due to various factors including policy preferences, party loyalty, or historical voting patterns among older demographics.

Conclusions

In this study, we set out to predict the overall popular vote in the 2025 Canadian federal election using regression models with post-stratification, drawing on data from the General Social Survey (GSS) and the Canadian Election Study (CES). Our hypothesis was that a combination of demographic, socioeconomic, and geographic factors could forecast the majority party choice in terms of the popular vote. The methodology involved a multinomial logistic regression model to analyze the survey data, followed by post-stratification to adjust the sample to better reflect the population demographics.

Key results from the regression analysis indicated that age and education level are significant predictors of voting behavior, with older voters tending to favor the Liberal Party and Conservative Party and younger voters inclined towards the NDP. The post-stratification results suggest that the Conservative Party could receive the largest share of the popular vote, with the Liberal Party and the NDP also maintaining significant support.

Looking at the big picture, these findings provide valuable insights into the Canadian political landscape and the factors that may influence the next federal election. However, the study has limitations, including the assumption of homogeneity within provinces and education levels due to the use of averages for Figure 6, which made the computation more manageable but at the cost of accuracy.

For future work, More detailed statistics that can take into consideration variations among groups should be included; this is especially important in provinces with varied populations. The accuracy of the model might also be improved by looking at additional possible predictors, such as economic factors or particular policy concerns.

In conclusion, this study advances knowledge of Canadian political dynamics and shows how statistical models may be used to forecast election results. Despite these drawbacks, this research provides a baseline for next electoral prediction models and lays the groundwork for more thorough examinations as further data becomes accessible and the 2025 election draws near. The findings emphasize the need of taking a variety of factors into account when projecting election outcomes and the necessity of ongoing improvement of prediction approaches in politics

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)

Appendix

Generative AI Statement

Here is where you can explain your usage of Generative AI tool(s). Be sure to reference it. For instance, including something like:

I used the following generative artificial intelligence (AI) tool: ChatGPT (September 13 version) [4].

- For the Introduction and Data sections, I sought explanations of political terminologies and statistical methods to ensure clarity and accuracy in the descriptions provided.
- In the Methods section, I requested that ChatGPT detail the multinomial logistic regression and post-stratification approach, ensuring the methodology was correctly applied and described. Since I did not know how to do multinomial logic regression, I asked it to write the R code for me and I copy-pasted it. I double-checked the results at each step with ChatGPT to make sure I was following the correct steps.
- I discussed the implications of using averages for province and education levels in the model with ChatGPT, prompting a concise explanation of the assumptions involved and their impact on the study's results.
- Finally, in the Conclusions section, I asked ChatGPT to summarize the study, emphasizing key findings, acknowledging weaknesses, suggesting future work, and crafting a concluding paragraph to encapsulate the study's contributions.

Each prompt was followed by a refinement or an additional request for more detailed explanations as needed.