A Study on Predicting the Canadian Federal Election in 2025

Group 123: Lulu Sheng, Shuqi Chen, Shutong Chen, Yuxin Zhang

November 24, 2022

Introduction

A healthy democracy requires the active participation of all citizens in political activities, and the general election is a concentrated expression of the political behaviour of all citizens. In Canadian federal elections, voters elect the Prime Minister indirectly by voting for their local Member of Parliament (MP) (Elections Canada). In other words, federal elections in Canada are more about voting for the parties. From the perspective of Canadian citizens, the outcome will determine the path of the country, which is meaningful to Canadian citizens. In terms of the parties running for election, they need to know what areas they should improve to get a higher number of votes, and this is important to 5 parties themselves.

In fact, election results are influenced by many personal factors. For example, age and gender affect voter results. There are gender differences in voter results, with younger generations more likely to vote for Democrats in terms of the age difference between voters of different parties. In terms of gender differences in voter results, there are more female voters than male voters who vote for the Democratic Party. In terms of religious differences, the difference in people's religious beliefs is reflected in political orientation. In the U.S. election, more and more non-religious voters voted for the Democratic Party. (Atske, 2020) In addition, voters' income level, place of birth and residence also influence their preferences for different political parties. For example, citizens who pursue radical economic policies will vote for the party that matches their desired policies (Anderson et al., 2004). And voters in some regions will vote for the party in that region (Kang, 2016). In addition, one's political perceptions change with the status of employment (Hackenberger et al., 2021).

The research question of this report is which party will receive the most votes. We assumed that age, gender, province of residence, home income, religion and employment status would be valid predictors of the outcome of the Canadian federal election. The selected predictors are used to make a logistic regression model to come up with the most reasonable model that predicts the number of votes the 5 parties will receive in 2025. Moreover, the 5 parties that we want to predict the voting results by logistic regression model are Liberal, Conservative, New Democratic Party, Bloc Québécois, and Green Party, which are all favourable contenders because they have the highest vote share in the previous years' general elections (Heard). Based on the votes received in previous years, the main hypothesis is that in 2025, the Liberal will have the highest share of votes received.

We used a total of two datasets. Firstly, General Social Survey (GSS) - Census Data was conducted in Canada in 2017 and was a sample phone survey. In addition, CES-survey data is the data we use to model the attitudes and opinions of Canadians (what parties they support) during and after the 2019 federal election using a phone survey. During the research, we used the logistic regression method based on past results modelling (CES survey) and then introduced data from the GSS census to predict the future dependent variable. We assume that all these 6 predictors mentioned before have an impact on voter results. These predictors will be studied specifically in the report.

Data part

Data Collection

CES (2019 Canadian Election Survey):

CES refers to the Canadian election survey used to build a model for the survey dataset. The main objective of this survey is to ensure that Canadians who are entitled to vote are able to use their democratic right to vote and become candidates. This survey is mainly used to collect information by telephone and mail, and the phone 2019 survey is used in this model (Election Canada). We grab this data from CES official website as our survey dataset.

GSS (2017 Canadian General Social Survey):

GSS is the primary source of census data referring to the Canadian general social survey. The general survey is to understand the sense of identity and belonging of Canadians to their society, culture and environment. The target population is Canadians aged 15 and over who have lived in 10 different Canadian provinces for 10 years. And this data was collected through online questionnaires and telephone interviews (Government of Canada, 2021). We grab this data from CHASS official website as our census dataset.

Data cleaning process

For the cleaning process of the census data, we first created a new variable representing participants' ages in the year 2019 by adding two to their age in the year 2017 to match the survey data collected in 2019 and round the ages to integers. Then converted the variable type of their sex, province living currently, family income, work and religion participation from vector to factor to make them category variables. We treated people choosing options non-response and refusal as missing values. Then we selected six variables: age, sex, province, family income, work and religion participation.

Regarding the cleaning process of the survey data, we first created a new variable indicating respondents' ages using 2019 to subtract their born years. Then we rename the variables in sex, provinces, family income, work, and religion participation to match the variable names in census data. Also, we converted variables from vectors to factors. We would not consider participants selected as "others" in sex or respondents who live in Yukon, Northwest Territories, and Nunavut because census data does not contain that information. We treated them as missing values. In addition, variables state whether the participants intend to vote for particular parties are created including Liberal, Conservatives, NDP, Bloc and Green Party. Also, we created a new categorical variable indicating participants' vote intention. Next, we selected twelve variables: age, sex, vote Liberal, vote Conservatives, vote NDP, vote Bloc, vote Green Party, vote intention, province, family income, work, and religion participation.

At last, we deleted all missing values in both datasets, created a unique identifier "ID" column for all cases, and made our data clear.

Variables introduction

Brief description:

- **Age (Numerical)** refers to Individual age.
- **Sex (Categorical)** is a binary variable indicating males or females.
- **Province (Categorical)** is the territory of current residence in Canada including 10 levels.
- **Family income (Categorical)** is the total household income with 6 levels before tax in 2018.
- **Work status (Categorical)** is the status Canadian citizens work or not.
- **Religion participation (Categorical)** refers to how many times people will be engaged in religious activities.
- **Vote intention (Categorical)** is the voting opinions of people who fill out the survey.

- **Vote Liberal (Categorical)** is whether participants vote the Liberal Party or not.
- **Vote Conservatives (Categorical)** is whether participants vote the Conservative Party or not.
- **Vote NDP (Categorical)** is whether participants vote the New Democratic Party or not.
- **Vote Bloc (Categorical)** is whether participants vote the Bloc Québécois or not.
- **Vote Green (Categorical)** is whether participants vote the Green Party or not.

Description for important variables

There are four essential predictors will be used in our model: sex, age, province and religious participation. The proportion of male and female voters affects the outcome of each political party's vote. We mentioned in the introduction that male and female have different party preferences. For example, more female voted for New Democracy than male age also affects preferences, with the younger and older groups having different views on political parties. Voters from their provinces are likely to account for the majority of the votes received by the party leaders. Also, the views of each party on religion and the religious involvement of their leaders can influence the decision of voters.

Numerical Summary

Table 1: Summary table for age in census and survey data

Variable	Minimum	Q1	Median	Q3	Maximum	IQR	Mean	Stand deviation
age(survey)	18	39	53	66	93	27	52.08541	17.06337
age(census)	18	40	56	69	82	29	54.26510	17.66770

Table 1 shows the data analysis about age in the survey and census data. In survey data, the range of age is from 18 to 93, and the median age is 53. In census data, the range of age is from 18 to 82, and the median age is 56. There is little difference in the standard deviation of age between the two data (Ces is 17.667696, Svy is 17.0633742), with the average age fluctuating around 53 years (Ces is 54.2650983, Svy is 52.0854054).

Table 2: Summary table for sex in census and survey data

Gender	Sample Count	Sample Proportion	Census Count	Census Proportion
Female	744	0.4021622	11010	0.5441336
Male	1106	0.5978378	9224	0.4558664

Table 2 shows the number and proportion of genders in both datasets. First, the sample sizes of the two data sets are different. In the survey data, the proportion of males (0.5978378) is almost 20% more than that of females (0.4021622). However, in the census data, the opposite result is observed, and the proportion of females (0.5441336) is almost 10% more than that of males (0.4558664).

Table 3: Summary table for province in census and survey data

Province	Sample Count	Sample Proportion	Census Count	Census Proportion
Alberta	140	0.0756757	1693	0.0836710
British Columbia	377	0.2037838	2466	0.1218741
Manitoba	140	0.0756757	1169	0.0577740
New Brunswick	83	0.0448649	1317	0.0650885
Newfoundland and Labrador	75	0.0405405	1080	0.0533755
Nova Scotia	95	0.0513514	1409	0.0696353
Ontario	381	0.2059459	5492	0.2714243
Prince Edward Island	99	0.0535135	692	0.0341999
Quebec	332	0.1794595	3784	0.1870120
Saskatchewan	128	0.0691892	1132	0.0559454

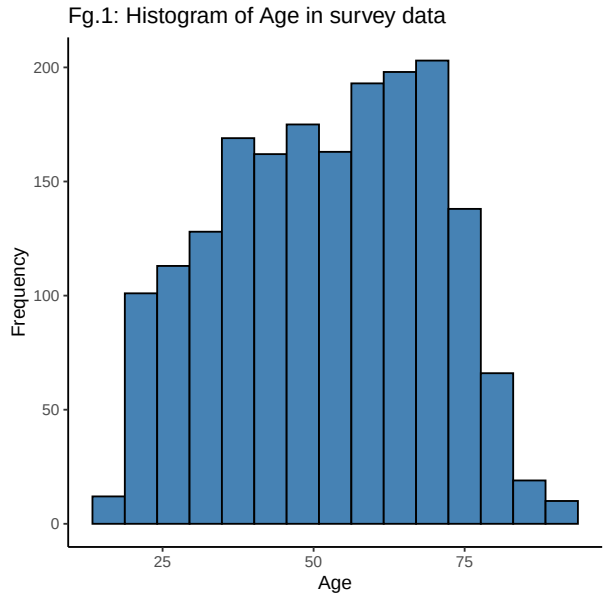
Table 3 shows the number and proportion of provinces in the two datasets. Three provinces stand out in both datasets: British Columbia(BC), Ontario(ON), and Quebec(QC). In the survey data, BC (0.2037838) and ON (0.2059459) are very close to each other, with QC (0.1794595) being slightly smaller. However, in the census data, ON (0.2714243) is more than the other two provinces, followed by QC (0.187012) and BC (0.1218741).

Table 4: Summary table for religion participation in census and survey data

Religion participation	Sample Count	Sample Proportion	Census Count	Census Proportion
At least 3 times a year	289	0.1562162	2125	0.1050213
At least once a month	183	0.0989189	1845	0.0911832
At least once a week	219	0.1183784	3360	0.1660571
Not at all	891	0.4816216	9783	0.4834931
Once or twice a year	268	0.1448649	3121	0.1542453

Table 4 shows the number and proportion of religious participation categories in both datasets. It is very clear to see that the proportion of voters with no religious activity is almost half in both datasets. Excluding the voters who do not have any religious involvement, we divide the remaining people into frequent (At least once a month + At least once a week) and occasional (At least 3 times a year + Once or twice a year). In the survey data, the proportion of voters who attend religious services occasionally (0.3010811) is higher compared to frequently (0.2172973). In the census data, however, the proportions are similar in both frequent (0.2592666) and occasional (0.2572403).

Graphical Summary:



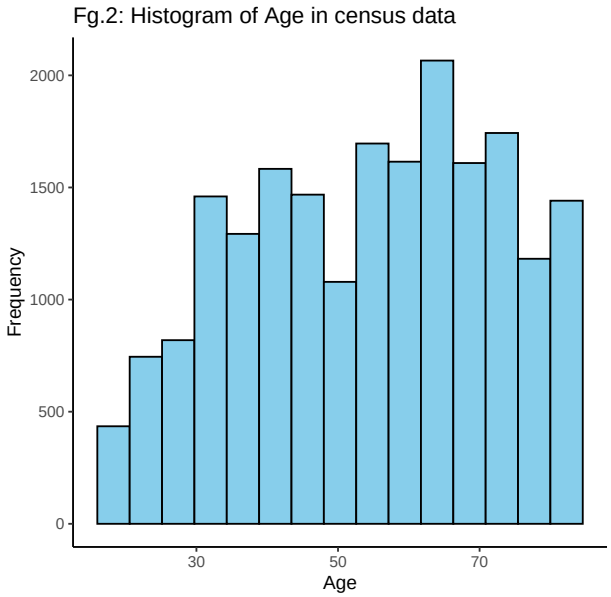
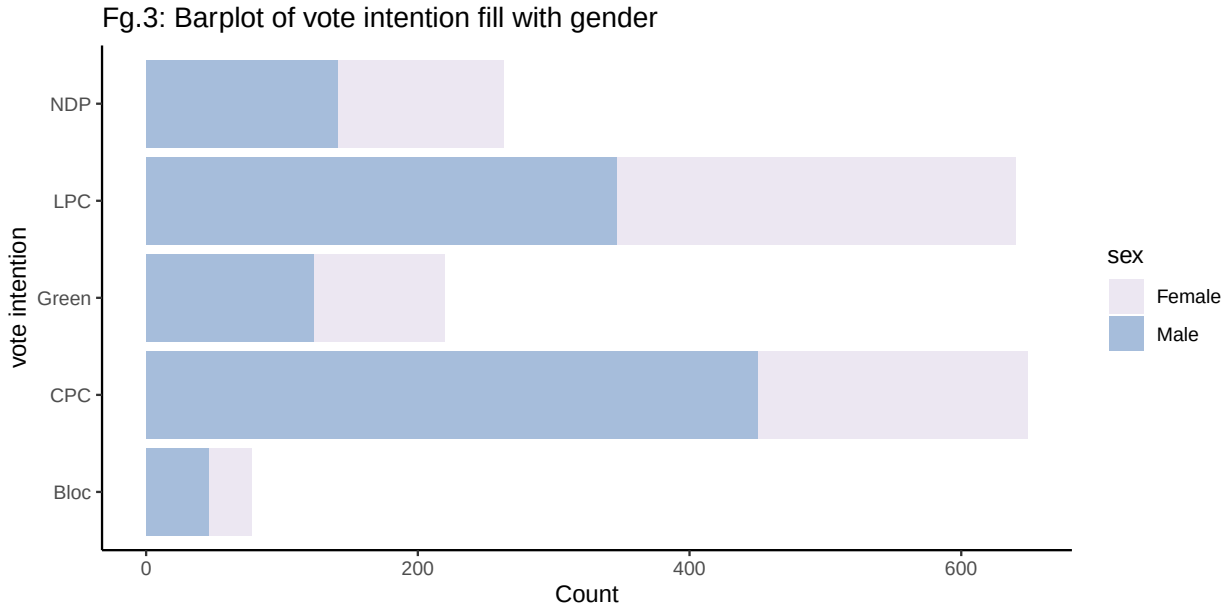
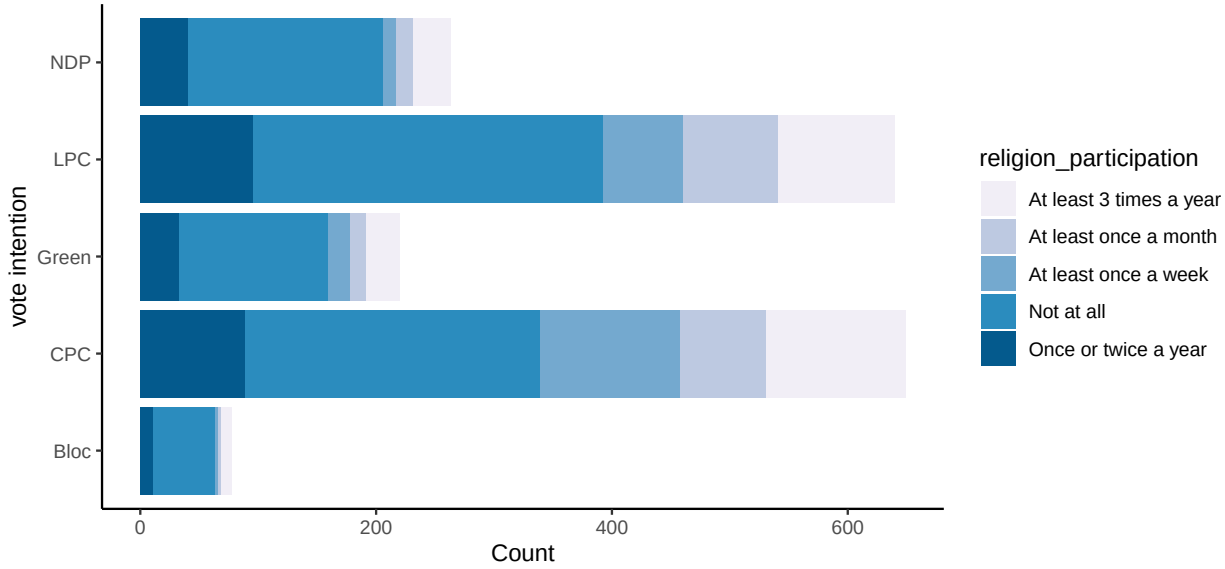


Figure 1 and Figure 2 show two histograms of age in the survey and census data separately. In the survey data (Fg.1), the age of most voters settled between 25 to 75 and the distribution of the histogram is normal. In the census data (Fg.2), the age of most voters settled between 40 to 80, and the distribution of the histogram is left-skew.



The bar chart (Fg.3) shows the number of votes received by the five political parties in the survey data and is coloured by gender. Conservatives and liberals received the most votes, and male voters made up the bulk of the conservative vote. There is little difference in the gender ratio of liberal voters.

Fg.4: Barplot of vote intention fill with religion participation



The bar chart (Fg.4) shows the five parties' voting results coloured by religious participation. We focus on conservatives and liberals. In contrast, voters who participate in religious activities once a week are more likely to vote for conservatives.

Methods

Model Selection

Logistic regression is a statistical model that uses a logistic function to predictor a binary response variable. In our study, we would build five logistic regression models to predict whether the participant intends to vote for each party. We selected six potential variables for building models previously. However, it is not necessary to contain all predictors because the too complicated model may cause overfitting problems leading to the coefficients of regression models and R squared being misleading (Minitab Blog, 2022). Thus, we decided to use Akaike Information Criterion (AIC) to assistant finding the most reasonable models for each party. AIC is a mathematical method to determine how well the model fits the data. The model with the lowest AIC among all the possible models containing six potential predictors could be considered the best model under the AIC criteria. We applied forwards stepwise selection (step function in R), which is a variable selection method beginning with the model without variables. The AIC of the model decreases by adding necessary variables and the adding process continues until AIC would not reduce by adding any predictors.

The final models used to predict citizens in census data intend to vote liberal or conservatives or green party consist of the same four predictors: age, sex, province and religion participation. We decided to use this final model in all parties. Although the selected models of bloc and NDP are different, we tried to use AIC function to compare the AIC of the NDP's selected model and the AIC of the final model of NDP: 1379.231 and 1384.596, which are almost identical. The AIC of the bloc's selected model and the AIC of the final model applying in the bloc are 374.609 and 380.528, which are also similar. Thus, we would use the same four variables to predict the election result in 2025.

Model Specifics

We will use a logistic regression model to model the proportion of voters who will vote for each party (five different models within the same predictors). The logistic regression model we are using is:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 age + \beta_2 sex + \beta_{3,i} province + \beta_{4,j} religion$$

p refers to the probability of the participant voting for a particular party.

β_0 refers to the log odds if they meet all the base levels and 0 years of age. It does not make sense because we only consider participants at least 18 years old in the study.

β_1 represents the change in log odds $\log\left(\frac{p}{1-p}\right)$ for every one unit increase in age.

β_2 represents the difference between the average log odds of female and male groups with fixed age, province and religion.

$\beta_{3,i}$ includes nine betas for $i = 1, 2, ..., 9$ refer to ten provinces. $\beta_{3,i}$ indicates the differences between the average log odds of Alberta (base level) and other provinces' when age, sex and religion are fixed.

$\beta_{4,j}$ consists of five betas for $j = 1, 2, ..., 4$ corresponding to five religious participate frequency groups. $\beta_{4,j}$ represents the difference between the mean log odds of the base level religion participation and other participate levels when age, sex and province are fixed.

Logistic regression assumption

As we are using a logistic regression model to predict which party will receive the most support, four assumptions need to be checked before using this method. These four assumptions are: a binary outcome, linearity of the log odds of the numerical predictors, multicollinearity and no significant outliers. First, let's check the assumption of a binary outcome. From the bar chart above, we can see that there are two categorical response variables for voting. That is, whether Canadian citizens would vote or not, which suggests that our results are binary. Next, let us check the linearity for numeric predictors which is 'age' in the model. In other words, we need to check $\log(\frac{p}{1-p})$ with repect to age. However, we only have one numeric variable - age, which is obviously linear from the histogram in the graphical summary. For the third assumption check, multicollinearity implies the presence of dependent predictors in the model. In this case, we use R function vif() to calculate the variance inflation factors. We can then obtain results for all predictors that are close to 1. Since all VIF values does not exceed 5, we can conclude that predictors are correlated with each other. Finally, we need to check the influential values in the model. We use cook's distance to visualize the plot for 5 groups, and then we find 4 largest outliers in each model. For example, in the liberal model, we have 123, 246, 998, 1012 four influential values. However, when checking these rows we find that these values are not weird. The reason they are considered as a outlier is because people who have a younger age but a higher income. For example, one male aged 36 but with an income over $125,000 in a year. However, this is a reasonable situation and we will not consider to remove these outliers.

Post-Stratification

Postratification formula:

$$\hat{y}^{PS} = \frac{\sum N_j \hat{y}_j}{\sum N_j}$$

In order to estimate the proportion of voters in 5 different parties, post_stratification method is introduced in this section. At first, based on the model we found in the logistic regression section above, the proportions of voters of different ages, genders, provinces and religious participation in our survey dataset will be applied to predict the most popular party in the next Canadian federal election. Having obtained the estimated proportions for the five parties, we then need to find a weighted estimate of the proportion of each party and the total population size.

Since we have 6 predictors, and each has a different number of variables. Then we multiply the variables together to get the total number of cells.For example, in the voting Liberal group, we have a total of 4395 cells. After complementing these preparations, we will use the formula $\hat{y}^{PS} = \frac{\sum N_j \hat{y}_j}{\sum N_j}$ to find the poststatification values. Here, j refers to 5 different groups, and $\hat{y_j}$ represents an estimate of the proportion of voters for five political parties. Also,N_j represents the total population size in each group. The reason why we chose to use post-stratification is that our data is derived from telephone interviews or online surveys. In other words, we do not know in advance the personal information of the people filling out this survey, which makes it difficult for us to stratify. Therefore, we chose the post-stratification method, where we stratify the information after we have it (Glasgow, 2005).

Results

Table 5: Result of the Logistic Regression Models

	Liberal	Conservative	NDP	Bloc	Green
(Intercept)	-2.401076	0.485350	-0.872032	-22.782987	-2.741891
age	0.007509	0.012534	-0.034639	0.026307	-0.007886
Province:					
British Columbia	1.177650	-2.146322	1.268609	-0.186535	1.693017
Manitoba	1.233692	-1.438246	0.832615	-0.157260	0.779972
New Brunswick	1.466389	-2.026563	-0.222797	-0.079627	2.025134
Newfoundland and Labrador	2.058852	-2.893225	1.809756	-0.069621	-0.012789
Nova Scotia	1.686402	-2.336495	0.864588	0.056230	1.477903
Ontario	1.924720	-2.110549	0.639246	-0.118094	0.922362
Prince Edward Island	1.820981	-2.333904	-0.020105	-0.067508	1.799608
Quebec	1.682736	-2.690649	0.216559	20.367830	0.549068
Saskatchewan	0.567032	-0.641447	0.726398	0.019064	-0.336284
Religion participation:					
At least once a month	0.505848	-0.197641	-0.312222	-1.011936	-0.425146
At least once a week	-0.096364	0.445958	-0.766655	-0.403583	-0.213648
Not at all	0.079972	-0.647532	0.586120	-0.000294	0.367039
Once or twice a year	0.106496	-0.328886	0.360325	-0.547563	0.252331
sex Male	-0.315074	0.745858	-0.451667	0.101057	-0.213485

Table 5 shows the coefficients of logistic regression models for the five parties. The intercept represents the log odds of the voters in each party when the voter's sex, province and religious participation is the base level and the age is 0. It is clear that the coefficients are positive in all provinces of the Liberal Party model, but the Conservatives show the opposite result. It means when age, sex and religious participation are held, the change in the log odds of voters living in other provinces who intend to vote for the liberal party instead of voters living in Alberta is positive, and the change in the log odds is negative in the conservative model. 20.3678 indicates that when holding age, sex and religion, the increase of the log odds of Quebec voters will vote for the Bloc instead of voters living in Alberta is 20.3678. Voters in Quebec have an extremely high approval of the Bloc. The effect of voter age does not have a significant impact on voting propensity for each party, but among them, the NDP ($\beta_1 age$ is -0.0346) and Green ($\beta_1 age$ is -0.0079) show negative age coefficients, which means that more younger people vote the NDP and Green relatively. In terms of the sex coefficient, 0.7459 states that the change in the log odds of male voters wanting to vote for the Conservatives instead of female voters is 0.7459 if age, religion and province are fixed. Thus, males are more likely than females to vote for the Conservatives. 0.446 represents the increase in the log odds of people who attend religious activities at least once a week will vote for the Conservatives instead of people who attend at least 3 times a year is 0.446 with fixed age, sex and province. Lots of voters who regularly participate in religious activities intend to vote for the Conservatives. -0.6475 indicates that when age, sex, and province are held, the change of the log odds of voters who never attend religious activities will vote for the Conservatives instead of voters who attend at least 3 times a year is -0.6475. It means people who never participate in religious activities are likely not to vote for the Conservative Party.

Table 6: Prediction of 2025 election

Party	predict proportion
liberal	0.3724824
conservatives	0.3347595
NDP	0.1359924
Bloc	0.0473546
Green	0.1142916

In Table 6, the results of the post-stratification shows that the Liberal Party (0.3724824) is expected to receive the most votes, followed closely by the Conservatives (0.3347595), while Bloc (0.0473546) is expected to receive the least number of votes among the five parties.

It is not reasonable for the whole prediction since if we add five proportions up, we will get 1.0048805. The summation is larger than 1 which is not possible. Therefore, the outputs might be overestimated by the models. However, it is reasonable to predict the result that the liberal will win the Canadian election in 2025 because although its supporting rate might be lower than 0.3724824, the vote of others also overestimate. Therefore, perhaps the supporting rate of others are lower than the result as well.

Conclusions

Summary

In this study, we first made a hypothesis, based on some background research, that the liberal party would receive the most votes in the next Canadian federal election in 2025. We then used age, gender, household income, province and religion as predictors and five different parties as responses, and built five logistic regression models to estimate them. After obtaining the predicted proportion of votes for each of the five parties, we stratified the four predictors by post-stratification method, adjusted the exact weight of each predictor in the model, and finally weighted the predicted proportion of votes of the five parties in the total population.

The final results show that by our logistic regression model, the four predictors we examined do have an effect on vote share across parties. The impact of the proportion of voters for these four variables varies among voters of different parties. For example, young people are more likely to vote Liberal, people living in Quebec are more likely to vote Bloc. What's more, within NDP voters, more voters barely participate in religious activities. In terms of sex, male voters are more likely to vote Conservative.

One of the important results is the predict proportion of liberal party is 0.3725. It turns out there are approximately 37.25% in Canada will support the liberal party. And it proved that our hypothesis is reasonable and correct, that we expect Liberal to receive the highest number of votes among the five parties in 2025. It is followed closely by Conservatives with 33.48% numbers of votes obtained, and Bloc to receive the lowest number of votes with 4.74%.

Limitations

Although we have built the logistic regression model and calculated the predicted votes using poststratification, there are still several limitations to our study. Here, we will talk about 4 limitations in our model. First of all, our census data is from 2017, which is 5 years old. This dataset might not be an adequate representation of the current Canadian population. And it is actually somewhat unreasonable for us to use this data to predict the 2025 Canadian federal election. In addition, as so much has changed in recent years, it is possible that many voters have changed their minds and will vote for a different party next term. For example, the impact of COVID-19 caused a lot of controversy about whether to wear masks and vaccinations, and the attitudes and behaviours of different political parties will reap the benefits of Canadian citizens' support or opposition, which will affect the vote in the next term. BomSimt (n.d) states that conservatives may have the most support in the next federal election. This conclusion is clearly different from our results. The second point is that our prediction models are only based on 4 predictors, which means that our models are not accurate enough. Next, We did not break down the age groups into subgroups when we cleaned the data, but rather into groups of one year. It has caused us to over-fit our model and thus overestimates our results as well (Brownlee, 2020), which resulted in the final predicted percentage of votes adding up to more than 1, for a total of more than 0.0048. Finally, Since we deleted all the missing values in the clean data process, we will have very few variables. For example, in our survey data, we started with 4021 observations, but when we removed the missing values, we had 2630 observations. It is possible that in these missing values person only had one question that was not filled in, but we then removed his entire observation. To Sum up, outdated datasets, fewer predictors, over-fitting models, and missing values are the four limitations of our study.

Next Steps

In future analyses, other factors that can influence voting choices, such as marital status, education level and first language, will need to be considered. These potential factors may help the model to be more accurate. In addition, the model might be built using an updated and more comprehensive dataset, which would allow us to make more accurate predictions if newer census data and survey data were available. Overall, updated datasets and many other potential factors are the next steps we focus on.

Bibliography

1. Anderson, C. J., Mendes, S. M., & Tverdova, Y. V. (2004). *Endogenous economic voting: Evidence from the 1997 British election.* Electoral Studies, 23(4), 683–708. https://doi.org/10.1016/j.electstud.2003.10.001

2. Atske, S. (2020, June). *In Changing U.S. Electorate, Race and Education Remain Stark Dividing Lines. Pew Research Center - U.S. Politics & Policy.* Retrieved November 28, 2022, from https://www.pewresearch.org/politics/2020/06/02/in-changing-u-s-electorate-race-and-education-remain-stark-dividing-lines/

3. Brownlee, J. (2020). *Why aren't my results as good as I thought? you're probably overfitting.* MachineLearningMastery.com. Retrieved from: https://machinelearningmastery.com/arent-results-good-thought-youre-probably-overfitting/

4. Canada, E. (n.d.). *The Electoral System of Canada. – Elections Canada.* Retrieved November 30, 2022, from https://www.elections.ca/content.aspx?section=res&dir=ces&document=part1&lang=e

5. Canadian federal election, 2025 (BobSmit). Future. (n.d.). Retrieved December 1, 2022, from https://future.fandom.com/wiki/Canadian_Federal_Election,_2025_(BobSmit)

6. Glasgow, G. (2005). *Stratified Sampling Types. University of California.* Retrieved from: https://doi.org/10.1016/B0-12-369398-5/00066-9

7. Hackenberger, A., Rümmele, M., Schwerter, J., & Sturm, M. (2021). *Elections and unemployment benefits for families: Did the family benefit dispute affect election outcomes in Germany?* European Journal of Political Economy, 66, 101955. https://doi.org/10.1016/j.ejpoleco.2020.101955

8. Hao Zhu (2021). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax.* http://haozhu233.github.io/kableExtra/, https://github.com/haozhu233/kableExtra.

9. Heard, A. (n.d.). *Canadian election results by party 1867 to 2021. Canadian Election Results: 1867-2021.* Retrieved November 30, 2022, from https://www.sfu.ca/~aheard/elections/1867-present.html

10. JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2022). *rmarkdown: Dynamic Documents for R. R package version 2.16.* URL: https://rmarkdown.rstudio.com.

11. John Fox and Sanford Weisberg (2019). *An {R} Companion to Applied Regression, Third Edition.* Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

12. Kang, W. C. (2016). *Local economic voting and residence-based regionalism in South Korea: Evidence from the 2007 presidential election.* Journal of East Asian Studies, 16(3), 349–369. https://doi.org/10.1017/jea.2016.19

13. Mine Çetinkaya-Rundel, David Diez, Andrew Bray, Albert Y. Kim, Ben Baumer, Chester Ismay, Nick Paterno and Christopher Barr (2021). *openintro: Data Sets and Supplemental Functions from 'OpenIntro' Textbooks and Labs.* http://openintrostat.github.io/openintro/, https://github.com/OpenIntroStat/openintro/.

14. Minitab Blog. (n.d.). *The danger of overfitting regression models. Minitab Blog.* Retrieved from: https://blog.minitab.com/en/adventures-in-statistics-2/the-danger-of-overfitting-regression-models

15. Thomas Lin Pedersen (2022). *patchwork: The Composer of Plots.* https://patchwork.data-imaginist.com, https://github.com/thomasp85/patchwork.

16. 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 (URL: https://doi.org/10.21105/joss.01686).

17. Yihui Xie (2022). *knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.40.*