Prediction of the Voting Results of the 2025 Canadian Federal Election

STA304 - Fall 2023 -Assignment 2

GROUP 21: Hazel Deng, Hao Ni, Tingbo Yuan, Yuyao Zhang

November 23, 2023

Introduction

Predicting the outcome of the upcoming Canadian federal election tentatively set for 2025 is the main goal of our analysis, relying on data from the General Social Survey (GSS) and the 2021 Canadian Election Study (CES). CES conducts research during federal elections to gather data on voter attitudes, opinions, and behavior. Through surveys and questionnaires administered to voters, CES aims to understand preferences, beliefs, and perspectives on candidates and policies using methods like telephone, face-to-face, or online surveys[4]. GSS in Canada captures diverse aspects of Canadian society, including attitudes, values, and behaviors. Covering education, health, family, employment, and political participation, GSS collects data via face-to-face interviews to provide insights for researchers and policymakers[7].

Within the evolving tapestry of Canadian politics, deciphering the pulse of the electorate stands as a cornerstone for policymakers and analysts alike. As the nation navigates through socio-economic transitions and ideological shifts, predicting voting behavior emerges as a paramount challenge with far-reaching implications. By employing statistical techniques logistic regression model and post-stratification, this study endeavors to bridge the gap between data-driven insights and informed electoral projections. Logistic regression is used to explore relationships between variables, particularly when the outcome is binary (e.g., yes/no, success/failure) and post-stratification is used to adjust sample data to match census data[8]. In predicting election results, logistic regression can analyze support for candidates or parties among voters and identify key factors influencing voting behavior. Post-stratification assists in adjusting sample data to better represent the entire electorate, such as different age groups, geographic locations, or other demographic characteristics, thus enhancing the accuracy and reliability of election result predictions.

Canada operates under a parliamentary system with a multi-party competition. Federal elections in Canada typically occur every four years, and each province has its unique electoral systems, with candidates representing various parties such as the Liberal Party, Conservative Party, New Democratic Party, Green Party, among others. Recent election results have showcased the dynamic nature of the Canadian political landscape. For instance, the competition between the Liberal and Conservative Parties has been a focal point in past elections. Additionally, smaller parties like the New Democratic Party and Green Party have demonstrated influence in certain regions[6]. Understanding these changes in the Canadian political sphere and the competition between parties will be central to our analysis. Navigating through the multifaceted landscape of election predictions demands an understanding of niche political terms that intricately shape electoral narratives. For instance, concepts such as "gerrymandering," the deliberate redrawing of electoral boundaries for partisan advantage, or the intricate strategies of "swing states," pivotal in determining electoral outcomes, form essential elements within our analytical framework.

Our research question is that which party will win the 2025 Canadian federal election based on a synthesis of GSS census data and CES survey insights. **A preliminary hypothesis emerges, projecting the Liberal Party's ascension to victory** based on historical trends, prevalent sentiments, and nuanced political landscapes. In the 2019 and 2021 Canadian federal election, the Liberal Party led by Justin Trudeau

was the winner in two elections[9]. This hypothesis leans on past electoral trends, the party's historical performance, and current socio-political indicators. While hypotheses set the foundation for our predictive analysis, it is essential to emphasize that they serve as initial conjectures and are subject to validation through rigorous analytical scrutiny. The notion of the Liberal Party's potential triumph draws on historical election data, regional support trends, and evolving public perceptions, thereby laying a preliminary framework for our predictive modeling efforts. Also, based on the results of the 2019 and 2021 elections, the Liberal Party and the Conservative Party are always the two of the major political parties in Canada[1]. Therefore, our report will focus on the models of the Liberal and Conservative Parties to compare the final prediction.

Our report unfolds into key sections: an exploration of data sources and variables, a description of methodologies including logistic regression and post-stratification, presentation of results, and interpretation of conclusions. Focused on predicting the 2025 Canadian federal election's popular vote using GSS census data and CES survey insights, these sections provide a comprehensive analysis of electoral predictions.

Data

The data collection process for GSS involves a multistage probability sampling technique to ensure representation across demographics and geographic regions. A comprehensive questionnaire covering social, economic, and political aspects is administered through various methods like face-to-face interviews or online surveys, aiming to gather information about individuals' behaviors, attitudes, and socio-economic backgrounds[7]. Similarly, CES employs survey design focusing on election-related aspects, administering questionnaires to a representative sample of Canadians via different modes[4]. Both surveys undergo data validation and cleaning post-collection to ensure accuracy and reliability.

Given that the aim of our study is to predict election outcomes using census data, based on a model estimated from survey data, it is imperative to standardize the variable names across both datasets. After thoroughly examining all variables in the two datasets, we identified six meaningful variables that could be uniformly measured.

We began by cleansing the survey data provided (ces2021.RData). The first variable is 'age,' representing the age of the voters.

The second variable is 'feelings of life,' indicating whether individuals are satisfied with their lives. There are five possible outcomes: Very Satisfied, Fairly Satisfied, Not very Satisfied, Not at all Satisfied, and Invalid. Notably, the original variable was numerical, ranging from 1 to 5, corresponding to these five outcomes. We converted all data into textual form and decided to exclude all instances marked as Invalid.

The third variable is the voter's sex. 'Sex' is a binary variable, with only two outcomes: male and female. However, the survey dataset contains a 'gender' variable, which also includes non-binary results. Therefore, we converted the numerical outcomes of the gender variable into text: male (1), female (2), non-binary (3), and invalid (4). Since it is challenging to determine non-binary sex from other variables, and the census data only includes a sex variable (not gender), we chose to retain only male and female results, discarding all data where gender is non-binary or Invalid. This step may introduce some inaccuracies and limitations, but we believe it is the best approach for handling non-binary results in this dataset.

Fourthly, the numerical results for the voter's province variable were also replaced with text. For example, Alberta is represented by 1, British Columbia by 2, and so on, up to Yukon represented by 13. We also removed all invalid results.

Fifth, the survey dataset provides exact numerical values for family income, such as 75,000 per year. In contrast, the census dataset categorizes family income into six brackets: Less than 25,000; 25,000 to 49,999; 50,000 to 74,999; 75,000 to 99,999; 100,000 to 124,999; 125,000 and more. Therefore, we converted the survey data's family income into the same categorical data as in the census dataset and excluded invalid results.

Sixth, the voter's marital status has six outcomes. However, these differ from the expressions used in the census data. Thus, we matched the survey dataset's marital status outcomes with those of the census, converting numerical results into text: Married (1), Living common-law (2), Divorced (3), Separated (4), Widowed (5), Single, never married (6), and removed all invalid results.

Finally, we added five variables representing whether the voter supported the Liberal, Conservative, NDP, Bloc Québécois, or Green parties. The outcomes are 'yes' or 'no.' For instance, a person who voted for the Liberals would have 'yes' for the Liberal vote and 'no' for all others.

Subsequently, we processed the census dataset. For the age variable, since our survey data is from 2021 and the census data from 2017, we added four years to all ages in the census data to align with the survey data.

Secondly, the census data's 'feelings of life' scores range from 1 to 10. We categorized these into the same four groups as in the survey dataset: scores of 10 and 9 as Very Satisfied, 8 to 6 as Fairly Satisfied, 5 and 4 as Not very Satisfied, and 3 to 1 as Not at all Satisfied. This categorization, based on our interpretation of the textual meanings, might affect the accuracy of this variable. However, we decided to retain it due to its significance.

The other three variables were processed similarly to the survey dataset, removing all invalid data and any missing data related to the variables used in both the survey and census datasets."

Data Visualization

Table 1: Summary of the age in the survey data

Min	Q1	Median	Q3	Max	Mean
18	36	53	65	97	51.11526

Age is the only numerical variable among our predictors. **Table 1** shows the summary of age. The mean age is 51, the minimum age is 18 and the maximum age is 97. There about 75% of people who completed the survey were aged below 65.

(Table.5-Table.9 are attached in Appendix)

To explore more about the data, we compare each predictor in survey and census data. **Table.5** compares the sex distribution. It shows higher counts in the census for both females (11,029) and males (9,222) compared to the survey, where females are 7,590 and males are 6,482. This difference is likely due to the broader and more comprehensive population sample covered by the census. In both datasets, females are more represented than males, suggesting a consistent gender distribution across both survey and census.

Table.6 compares the feelings of life, it shows higher levels of dissatisfaction, with more respondents reporting being "Not at all Satisfied" or "Not very Satisfied," while the census data indicates a higher proportion of respondents feeling "Very Satisfied." This suggests that survey participants generally reported lower satisfaction levels compared to the broader population in the census.

Table.7 explores the living provinces, while some provinces like Alberta and Ontario have similar counts in both datasets, others like British Columbia, Manitoba, and Quebec show significant differences. Notably, the survey underrepresents provinces like British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, and Prince Edward Island compared to the census. Conversely, Quebec is over-represented in the survey. The territories (Northwest Territories, Nunavut, Yukon) have minimal or no representation in the census data, which contrasts sharply with their presence in the survey, albeit with small counts. This disparity suggests that the survey sample may not fully align with the actual provincial distribution of the population as captured in the census.

Table.8 indicates the family income. The survey underrepresents higher income categories ($125,000$ and more) and lower income categories (Less than $25,000$) compared to the census. Middle income ranges, particularly $25,000$ to $49,999$ and $50,000$ to $74,999$, show higher counts in the census than in the survey. These differences indicate a potential skew in the survey data towards middle and upper-middle-income ranges, while the census data presents a broader distribution across various income levels.

Table.9 is the comparison of marital status, while "Married" is the most common status in both datasets, the survey underrepresents this category (6,461 responses) compared to the census (9,374 responses). The

survey reports more individuals "Living common-law" (2,392) than the census (2,060), which is an inverse trend compared to other categories. For statuses like "Divorced," "Separated," "Single, never married," and "Widowed," the census consistently shows higher counts than the survey, indicating a broader representation of these marital statuses in the general population compared to the survey sample. Thus, by all tables above, there are significant differences between survey data and census data, that's why we should use census data to predict for the result.

Figure.1 Barplot of life feeling of those who voted the Liberal Party

Figure.2 Barplot of life feeling of those who voted the Conservative Party

We also use barplots to compare the distribution of life satisfaction of voters for Liberal and Conservative Parties. In both **Figure 1** and **Figure 2**, the horizontal axis represents four categories of life satisfaction: "Fairly Satisfied," "Not at all Satisfied," "Not very Satisfied," and "Very Satisfied," and the vertical axis represents the count of individuals within each category. For both of the parties, red color means the individual did not vote for this particular party, and teal color means the individual did vote for it.

The height of each bar indicates the number of individuals who fall into each category. It can be seen that the voters proportion for each party do not have large differences in most life satisfaction. However, for those who are very satisfied with their lives, there is a larger proportion that votes for Liberal party.

Based on **Figure.3** and **Figure.4**[13] which are attached in Appendix, it is evident that the age distribution in the survey data is more concentrated between 25 to 75 years old, with a notable decline in the number of participants above 75 years old. In contrast, the age distribution in the census data is more evenly spread across different age groups, with the highest frequency of responses observed around the age of 90 years old.

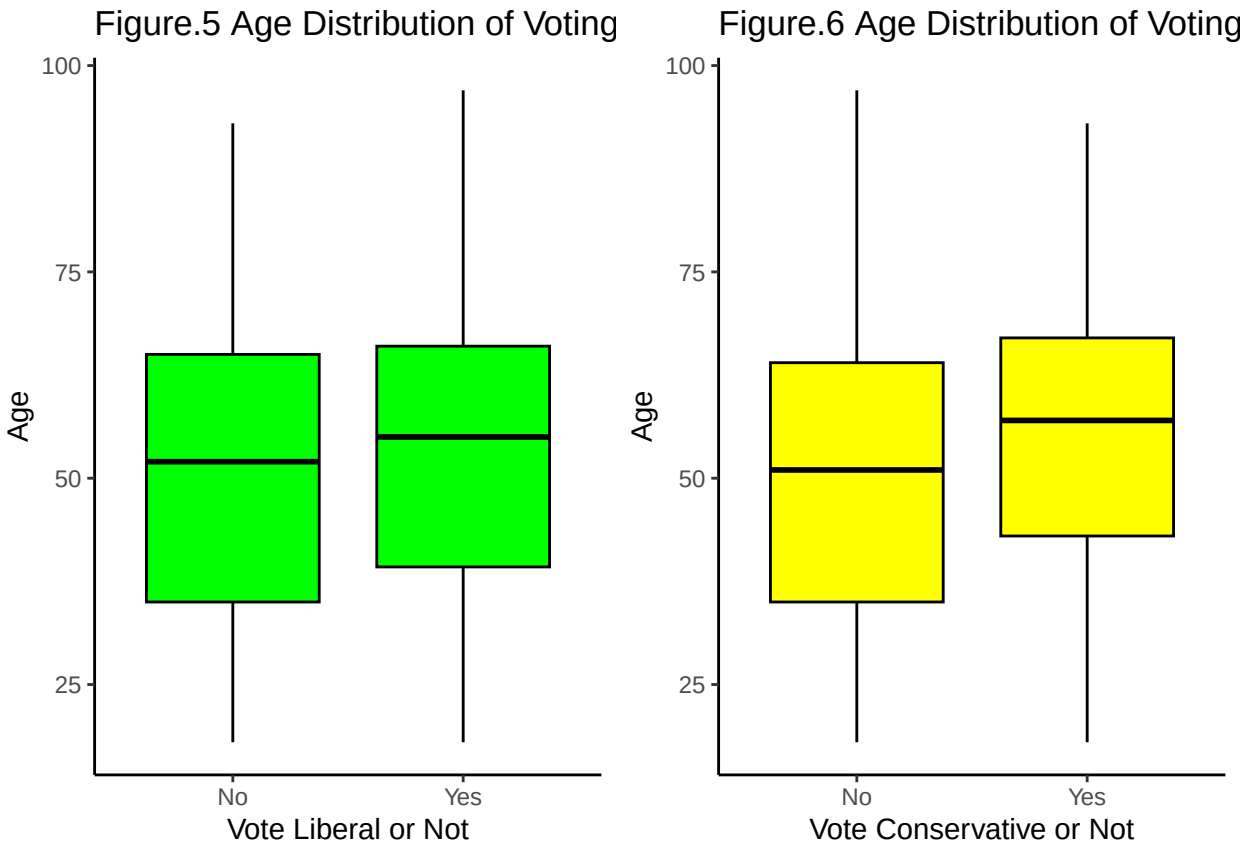

Figure.5 and **Figure.6** show the age distribution of voting two parties or not. Liberal voters and non-voters have similar age distributions, with a slightly younger median for voters. Conservative voters tend to be older, with a higher median age compared to non-voters. Both voters tend to be older than their respective non-voters. The interquartile ranges are quite similar across two parties, indicating no significant difference in the spread of ages among voters and non-voters.The presence of outliers is minimal, suggesting there are few very young or very old voters or non-voters across the parties.

Methods

Logistic regression model

The first step is to use logistic regression model to predict proportion of voters who will vote for the liberal party or conservative party. Logistic regression is a kind of predictive analysis that can be used when the response variable is binary. Based on the given data of some independent variables, logistic regression model can estimate the probability of an event occurring, such as vote or not[12]. In our project, logistic regression model is expected to predict the voting rate of different Canadian parties (especially Liberal and Conservative parties which are popular) in 2025 based on 2019 Canadian election study survey. The binary response is that if a person will vote one of the parties or not.

The six independent variables are the age, feelings of life, sex, provinces, family income, and marital status which have been described above. The age is as a numerical variable and all the remaining variables are categorical. Based on these six predictors, the logistic regression model is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{NAASatis} + \beta_3 x_{NotVerySatis} + \beta_4 x_{VerySatis} + \beta_5 x_{Male} + \beta_6 x_{inBC} + \beta_7 x_{inMN}$$

$$+\beta_8 x_{inNB} + \beta_9 x_{inNL} + \beta_{10} x_{inNT} + \beta_{11} x_{inNS} + \beta_{12} x_{inNU} + \beta_{13} x_{inON} + \beta_{14} x_{inPE} + \beta_{15} x_{inQB} + \beta_{16} x_{inSA}$$
$$+\beta_{17} x_{inYU} + \beta_{18} x_{income125+} + \beta_{19} x_{income25-49} + \beta_{20} x_{income50-74} + \beta_{21} x_{income75-99} + \beta_{22} x_{income25-} +$$
$$\beta_{23} x_{live} + \beta_{24} x_{married} + \beta_{25} x_{separated} + \beta_{26} x_{single} + \beta_{27} x_{widowed}$$

Where:

- p represents the probability of people will vote for a specific party.
- β_0 is the intercept of logistic regression model, and it is the log of odds of voting for each specific party when a newborn female infants from Alberta with income 100,000 - 124,999, and who are divorced and feel fairly satisfied.
- β_1 is the slope of age, and for every one-unit increase in age there will be a β_1 increase log odds of voting for a specific party.
- β_2 represents the coefficient of those are not at all satisfied with life.
- β_3 represents the coefficient of those are not very satisfied with life.
- β_4 represents the coefficient of those are very satisfied with life.
- β_5 represents the coefficient of those are male.
- β_6 represents the coefficient of those are in British Columbia.
- β_7 represents the coefficient of those are in Manitoba.
- β_8 represents the coefficient of those are in New Brunswick.
- β_9 represents the coefficient of those are in Newfoundland and Labrador.
- β_{10} represents the coefficient of those are in Northwest Territories.
- β_{11} represents the coefficient of those are in Nova Scotia.
- β_{12} represents the coefficient of those are in Nunavut.
- β_{13} represents the coefficient of those are in Ontario.
- β_{14} represents the coefficient of those are in Prince Edward Island.
- β_{15} represents the coefficient of those are in Quebec.
- β_{16} represents the coefficient of those are in Saskatchewan.
- β_{17} represents the coefficient of those are in Yukon.
- β_{18} represents the coefficient of those family income over $125,000.
- β_{19} represents the coefficient of those have $25,000 to $49,999 family income.
- β_{20} represents the coefficient of those have $50,000 to $74,999 family income.
- β_{21} represents the coefficient of those have $75,000 to $99,999 family income.
- β_{22} represents the coefficient of those family income less than $25,000.
- β_{23} represents the coefficient of those are living with common law.
- β_{34} represents the coefficient of those are married.
- β_{35} represents the coefficient of those are separated.

- β_{36} represents the coefficient of those are single and never married.
- β_{37} represents the coefficient of those are widowed.

The assumptions for logistic regression model should be checked:

1. Checking binary outcome: The dependent variable should be binary or dichotomous. This means it should have only two possible outcomes. Our binary response is that if a person will vote one of the parties or not.
2. Checking no multicollinearity: By looking at the Variance Inflation Factor (VIF), where a VIF value above 5 is often considered as an indicator of multicollinearity. Our VIF of all predictors are around 1. Therefore, there is no Multicollinearity among our independent variables.
3. Checking no strong outliers: by checking **Figure.7** and **Figure.8** in appendix, there are some outliers exist in two models. However, these are not influential outliers and we are expected to avoid removing some important data so keeping all the data are necessary.

Post-Stratification

Next, doing a post-stratification analysis after building each logistic regression model based on survey data. Post-stratification is a technique used in sample surveys to increase estimator efficiency[11]. We will create several cells with varying age, life feelings, sex, provinces, family income, and marital status. Then we use each logistic regression model to estimate the probability of voting for each Canadian political party in different cells, and we need to find weighted average of estimates from all possible combinations of attributes. The goal is to overcome inconsistencies in the characteristics of the participants in the survey and the population such as age distribution or distribution. It is also possible to improve the effect on estimates of groups with large population sizes. According to the post-stratification formula, we could change each cell with their respective population weight, then we sum them all and divide the total population size.

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

Where:

- $\hat{y}^{PS}$ represents the estimation after using the post-stratification method.
- N_j is the population size of each stratum j.
- N represents the total population size.
- $\widehat{y}_j$ is the sample mean for each stratum j.

Since the estimation is based on the survey data, there may be a bias here when the census data has various distribution on our predictors. It is crucial to use Post-Stratification analysis because post-stratification can reduce bias and improve the accuracy of survey estimates.

Results

Table 2: Coefficient of the model of Liberal Party

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.2782271	0.1370349	-9.3277501	0.0000000
Age	0.0070647	0.0014230	4.9645348	0.0000007
Life feeling: Not at all satisfied	-2.8421605	0.2280227	-12.4643731	0.0000000
Life feeling: Not very satisfied	-1.4599522	0.0721921	-20.2231571	0.0000000
Life feeling: Very satisfied	0.9992243	0.0510153	19.5867567	0.0000000
Male	-0.0430136	0.0423353	-1.0160220	0.3096189
Live in British Columbia	0.1587400	0.0894804	1.7740196	0.0760599
Live in Manitoba	0.1000635	0.1252879	0.7986690	0.4244823
Live in New Brunswick	0.6381950	0.1535942	4.1550736	0.0000325
Live in Newfoundland and Labrador	0.8523873	0.2009853	4.2410425	0.0000222
Live in Northwest Territories	0.7538625	0.7420352	1.0159390	0.3096585
Live in Nova Scotia	0.6533772	0.1384443	4.7194227	0.0000024
Live in Nunavut	-10.9467826	93.9503149	-0.1165167	0.9072430
Live in Ontario	0.4959541	0.0721530	6.8736459	0.0000000
Live in Prince Edward Island	0.4116157	0.4013424	1.0255974	0.3050814
Live in Quebec	-0.0674398	0.0762994	-0.8838841	0.3767588
Live in Saskatchewan	-0.5688159	0.1891300	-3.0075391	0.0026337
Live in Yukon	-0.2256466	0.5706467	-0.3954227	0.6925310
Family income: $125,000+	0.0536196	0.0727599	0.7369386	0.4611597
Family income: $25,000 to $49,999	-0.2449324	0.0812907	-3.0130424	0.0025864
Family income: $50,000 to $74,000	-0.0973221	0.0744992	-1.3063516	0.1914330
Family income: $75,000 to $99,999	-0.0024527	0.0758405	-0.0323397	0.9742012
Family income: $25,000-	-0.2414471	0.0835780	-2.8888842	0.0038661
Living common-law	-0.2246801	0.0947727	-2.3707255	0.0177532
Married	-0.1077957	0.0810926	-1.3292924	0.1837515
Separated	-0.0258636	0.1401683	-0.1845186	0.8536067
Single	0.0079549	0.0899445	0.0884425	0.9295250
Widowed	-0.0135393	0.1278599	-0.1058919	0.9156681

Table 3: Coefficient of the model of Conservative Party

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.7113861	0.1360802	-12.5763030	0.0000000
Age	0.0149558	0.0014593	10.2489182	0.0000000
Life feeling: Not at all satisfied	0.8973604	0.0774049	11.5930742	0.0000000
Life feeling: Not very satisfied	0.7034390	0.0501722	14.0204885	0.0000000
Life feeling: Very satisfied	-0.4797623	0.0663298	-7.2329867	0.0000000
Male	0.4188363	0.0425749	9.8376267	0.0000000
Live in British Columbia	-0.6598460	0.0808624	-8.1601053	0.0000000
Live in Manitoba	-0.3859125	0.1105840	-3.4897670	0.0004834
Live in New Brunswick	-1.0679260	0.1741246	-6.1331154	0.0000000
Live in Newfoundland and Labrador	-0.8878962	0.2315250	-3.8349912	0.0001256
Live in Northwest Territories	-0.4476091	0.7459237	-0.6000736	0.5484572
Live in Nova Scotia	-0.8877627	0.1489812	-5.9588906	0.0000000
Live in Nunavut	-10.6520192	95.8076191	-0.1111813	0.9114726
Live in Ontario	-0.4785059	0.0616786	-7.7580499	0.0000000

	Estimate	Std. Error	z value	Pr(>|z|)
Live in Prince Edward Island	-0.3670601	0.3605764	-1.0179817	0.3086867
Live in Quebec	-1.1586208	0.0691324	-16.7594449	0.0000000
Live in Saskatchewan	0.1883868	0.1333503	1.4127212	0.1577377
Live in Yukon	-0.4196976	0.5076277	-0.8267824	0.4083604
Family income: $125,000+	-0.0294399	0.0716742	-0.4107453	0.6812593
Family income: $25,000 to $49,999	-0.3688243	0.0807136	-4.5695445	0.0000049
Family income: $50,000 to $74,000	-0.2000305	0.0737428	-2.7125429	0.0066769
Family income: $75,000 to $99,999	-0.1636963	0.0750910	-2.1799722	0.0292595
Family income: $25,000-	-0.4822534	0.0840531	-5.7374860	0.0000000
Living common-law	0.1050627	0.0996998	1.0537906	0.2919788
Married	0.4551704	0.0847977	5.3677232	0.0000001
Separated	0.2195360	0.1442122	1.5223128	0.1279307
Single	0.1525314	0.0954992	1.5972018	0.1102207
Widowed	0.2494964	0.1320211	1.8898232	0.0587816

Table.2 and **Table.3** show the coefficients of each predictor in the Liberal and Conservative models, includes estimates, standard error, statistic, and p-value.

Fittiing model for Liberal Party:

$$log(\frac{p}{1-p}) = -1.278227+0.007065x_{age}-2.842160x_{NAASatis}-1.459952x_{NotVerySatis}+0.9992x_{VerySatis}-0.043014x_{Male}$$
$$+0.158740x_{inBC}+0.100064x_{inMN}+0.638195x_{inNB}+0.852387x_{inNL}+0.753862x_{inNT}+0.653377x_{inNS}-10.946783x_{inNU}$$
$$+0.495954x_{inON}+0.411616x_{inPE}-0.067440x_{inQB}-0.568816x_{inSA}-0.225647x_{inYU}+0.053620x_{income125+}$$
$$-0.244932x_{income25-49}-0.097322x_{income50-74}-0.002453x_{income75-99}-0.241447x_{income25-}$$
$$-0.224680x_{live}-0.107796x_{married}-0.025864x_{separated}+0.007955x_{single}-0.013539x_{widowed}$$

Fitting model for Conservative Party:

$$log(\frac{p}{1-p}) = -1.711386+0.014956x_{age}+0.897360x_{NAASatis}+0.703439x_{NotVerySatis}-0.479762x_{VerySatis}$$
$$+0.418836x_{Male}-0.659846x_{inBC}-0.385913x_{inMN}-1.067926x_{inNB}-0.887896x_{inNL}-0.447609x_{inNT}-0.887763x_{inNS}$$
$$-10.652019x_{inNU}-0.478506x_{inON}-0.367060x_{inPE}-1.158621x_{inQB}+0.188387x_{inSA}-0.419698x_{inYU}$$
$$-0.029440x_{income125+}-0.368824x_{income25-49}-0.200031x_{income50-74}-0.163696x_{income75-99}-0.482253x_{income25-}$$
$$+0.105063x_{live}+0.455170x_{married}+0.219536x_{separated}+0.152531x_{single}+0.249496x_{widowed}$$

Based on the coefficients of each predictor, two regression models can be determined.Also, the p-value in the last column of the coefficient table shows that marital status has p-values larger than 0.05, which means that it is not likely contributing significantly to explain the variability of whether or not to vote for liberal or conservative party. However, the p-value of most of the predictors are smaller than 0.05 and even closely to 0, which means the model fits well and the result of our model is reasonable and valuable.

After that, based on two regression models, using post-stratification that we introduced before to predict the voting probability of the Liberal and Conservative party in the 2025 federal election. According to the formula that is introduced in the Method part, the final result can be determined. **Table.4** shows the final result of our prediction. The voting probability of the Liberal party is 39.55%, and the voting probability of the Conservative party is 22.05%.

Table 4: Estimation of Voting Probability

Liberal	Conservative
0.3955494	0.2205211

According to this probability, it is possible to predict that the Liberal Party will win the 2025 Canadian federal election in the end, which is consistent with what we initially predicted from past elections. This is also the same result as the precious two elections in 2019 and 2021, which were won by the Liberal Party, led by Justin Trudeau[9]. It also justifies our model for two parties are reasonable. However, the actual result will still be related to the social environment, policies, and institutions at that time. We can say that the Liberal party has a better chance of winning in the 2025 federal election, which is consistent with what we initially predicted from past elections.

Conclusions

This study aimed to predict the outcome of the 2025 Canadian federal election by utilizing logistic regression models and post-stratification techniques. The initial hypothesis suggested a Liberal party victory, drawing from historical data trends and insights derived from the GSS census data and CES survey.

Logistic regression models predicted a 39.55% probability for the Liberal party and a 22.05% probability for the Conservative party to win the 2025 federal election. Factors such as age, life satisfaction levels, and province of residence emerged as influential predictors of voting behaviors. The models demonstrated reasonable predictive capabilities, although certain predictors, notably marital status, showed non-significant contributions.

The study highlight the effectiveness of logistic regression and post-stratification in predicting election outcomes. These insights hold substantial value for policymakers, analysts, and campaign strategists in preparing for the forthcoming election. Limitations in the study included reliance on historical data, which might not fully capture evolving voter sentiments. Moreover, some predictors showed non-significant impacts, suggesting potential gaps in the model. Also, there are some influential outliers in two models that are not removed in order to avoid wiping out important data. This may lead to the inaccuracy of the estimate of our models.

Future analyses should consider integrating real-time sentiment analysis from social media platforms and expanding the array of predictors to capture the dynamic nature of voter behavior more comprehensively. Incorporating diverse variables and real-time data sources could enhance predictive accuracy in future election forecasting.

In conclusion, while the study's findings offer significant insights into potential election outcomes, the predictive models' limitations underscore the need for a more comprehensive and dynamic approach to accurately capture evolving voter behaviors in Canadian federal elections.

Bibliography

[1] 45th Canadian Federal Election. (2023). https://en.wikipedia.org/wiki/45th_Canadian_federal_election.

[2] Auguie, B. (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3.

[3] Allaire, J.J., et al. *References: Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/docs/.

[4] Canadian Election Study. (2021). http://www.ces-eec.ca/.

[5] Çetinkaya-Rundel, M. et al. (2021). openintro: Data Sets and Supplemental Functions from 'OpenIntro' Textbooks and Labs. http://openintrostat.github.io/openintro/, https://github.com/OpenIntroStat/openintro/.

[6] Elections Canada. (2021). https://www.elections.ca/home.aspx.

[7] General Social Survey. (2017). https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2013001-eng.htm.

[8] Grolemund, G. (2014, July 16) *Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/articles_intro.html.

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

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

[11] Valliant,R. (1992). Post-Stratification and Conditional Variance Estimation. https://www.bls.gov/osmr/research-papers/1993/pdf/st930500.pdf.

[12] What is Logistic Regression? (n.d.) https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/.

[13] Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis. Springer*. https://ggplot2.tidyverse.org/.

[14] Wickham, H., et al. (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).

[15] Xie,Y. (2023). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.43, <URL: https://yihui.org/knitr/>.

Generative AI Statement

We used the following generative artificial intelligence (AI) tool: ChatGPT Version 3.5[9]. We used the tool before the report started, we gave it the following prompt of `Which party is the winner in the 2019 and 2021 Canadian federal election?` and it tells us the Liberal Party led by Justin Trudeau won the most seats in two elections. And we used the tool in the Method section of this assignment and I gave it the following prompt of `How to judge the multicollinearity of one model based on VIF results?` and it tells me that a VIF of 1 means no correlation among predictors. If it is less than 5 then it is considered acceptable, and values above 10 means high multicollinearity.

Appendix

```
## Rows: 20,251
## Columns: 6
## $ age            <dbl> 57, 55, 68, 84, 32, 67, 63, 84, 68, 29, 20, 44, 61, 31,~
## $ sex            <fct> Female, Male, Female, Female, Male, Female, Female, Fem~
## $ province       <fct> Quebec, Manitoba, Ontario, Alberta, Quebec, Quebec, Nov~
## $ income_family  <fct> "$25,000 to $49,999", "$75,000 to $99,999", "$75,000 to~
## $ feelings_life  <fct> Fairly Satisfied, Very Satisfied, Fairly Satisfied, Ver~
## $ marital_status <fct> "Single, never married", "Married", "Married", "Married~
```

```
## Rows: 14,072
## Columns: 11
## $ age                 <dbl> 22, 28, 41, 63, 52, 66, 48, 65, 66, 54, 68, 31, 29~
## $ feelings_life       <fct> Fairly Satisfied, Fairly Satisfied, Fairly Satisfi~
## $ sex                 <fct> Female, Female, Female, Female, Female, Male, Fema~
## $ province            <fct> British Columbia, British Columbia, Quebec, Quebec~
## $ income_family       <chr> "$125,000 and more", "$125,000 and more", "$75,000~
## $ marital_status      <chr> "Single, never married", "Married", "Married", "Ma~
## $ vote_liberal        <fct> No, No, No, No, No, No, No, No, No, Yes, No, No, N~
## $ vote_conservative   <fct> No, No, No, No, No, Yes, No, Yes, No, No, No, No, ~
## $ vote_ndp            <fct> Yes, No, Yes, No, Yes, No, Yes, No, No, No, No, Ye~
## $ vote_Bloc_Québécois <fct> No, No, No, Yes, No, No, No, No, No, No, No, No, N~
## $ vote_green          <fct> No, No, No, No, No, No, No, No, No, No, No, No, No~
```

These are the cleaned census dataset and survey dataset that are used in this report.

Figure.3 Distribution of age of people voting of survey data

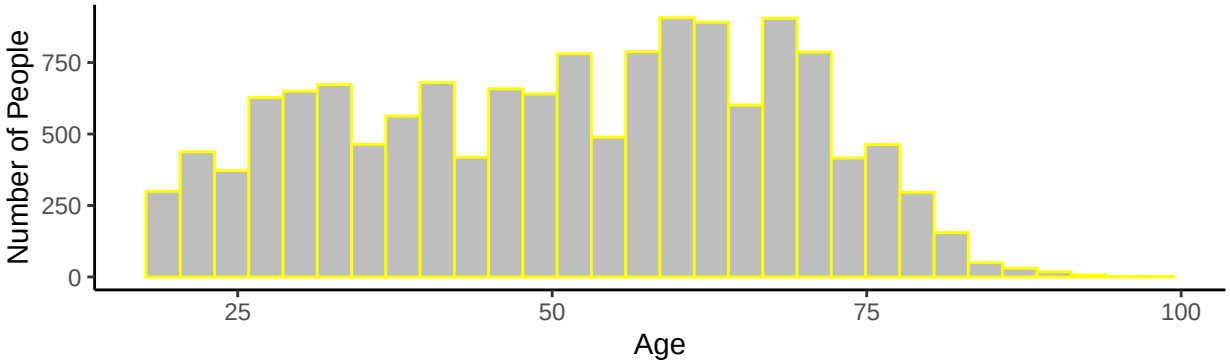

Figure.4 Distribution of age of people voting of census data

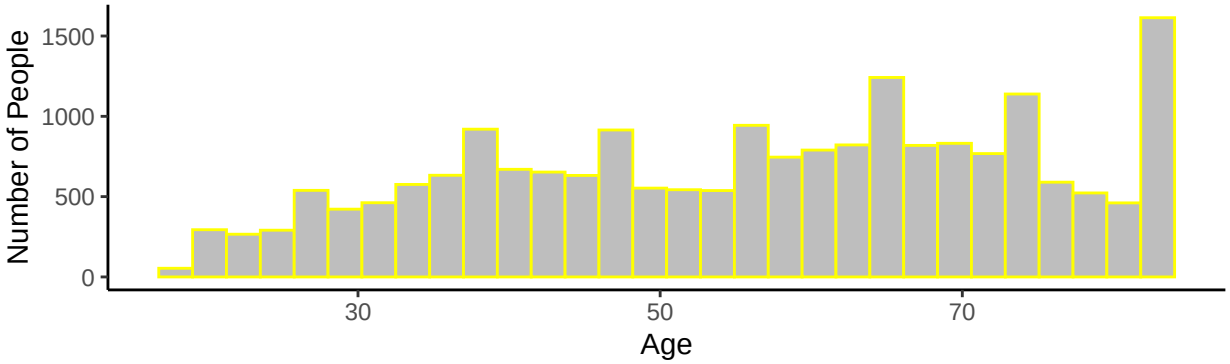

Table 5: Comparison of Sex Distribution in Survey and Census Data

Sex	Survey Count	Census Count
Female	7590	11029
Male	6482	9222

Table 6: Comparison of Feelings about Life in Survey and Census Data

Feelings about Life	Survey Count	Census Count
Fairly Satisfied	8327	10346
Not at all Satisfied	862	245
Not very Satisfied	2774	1250
Very Satisfied	2109	8410

Table 7: Comparison of Province Distribution in Survey and Census Data

Province	Survey Count	Census Count
Alberta	1680	1704
British Columbia	1511	2482

Province	Survey Count	Census Count
Manitoba	533	1166
New Brunswick	262	1308
Newfoundland and Labrador	130	1080
Northwest Territories	9	0
Nova Scotia	341	1396
Nunavut	4	0
Ontario	4993	5512
Prince Edward Island	38	699
Quebec	4256	3773
Saskatchewan	295	1131
Yukon	20	0

Table 8: Comparison of Income Family Distribution in Survey and Census Data

Income Family	Survey Count	Census Count
$100,000 to $ 124,999	1696	2137
$125,000 and more	2826	4661
$25,000 to $49,999	2282	4238
$50,000 to $74,999	2825	3645
$75,000 to $99,999	2403	2892
Less than $25,000	2040	2678

Table 9: Comparison of Marital Status Distribution in Survey and Census Data

Marital Status	Survey Count	Census Count
Divorced	1083	1722
Living common-law	2392	2060
Married	6461	9374
Separated	406	617
Single, never married	3237	4637
Widowed	493	1841

Figure.7 Cook's distance model of the Liberal Party

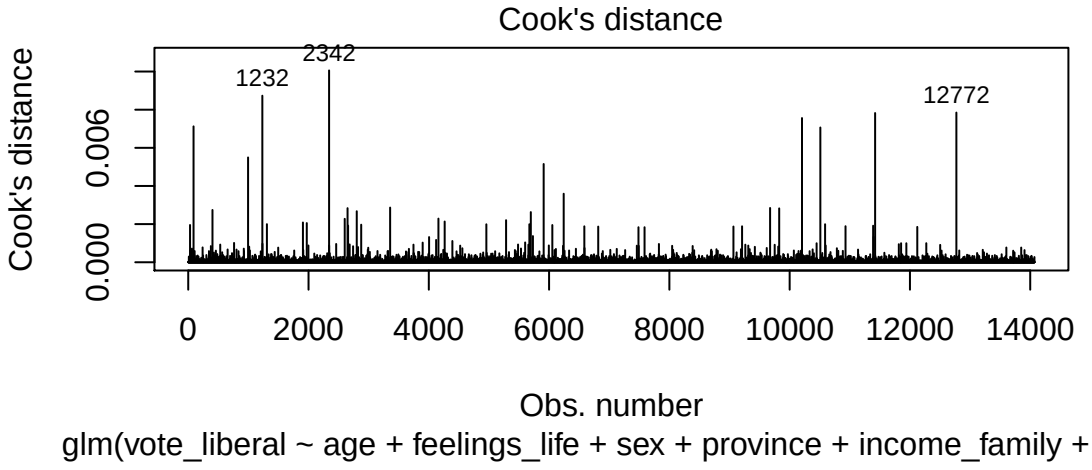

Figure.8 Cook's distance model of the Conservative Party