Analysis on Predicting Next Election Voting Results with Demographic Characteristics

STA304 - Assignment 2

GROUP 15: Aliza Aziz Lakho, Ting Jui Peng, Daniella Tong, Tsun Ting (Ryan) Tsai

November 24, 2022

Introduction

First of all, we have studied the "census" data from the General Social Survey (GSS), and "survey" data from the CES2019 package. We have brainstormed ideas about how we should pick the variables for our model. However, there are too many variables available in the dataset. Therefore, we decided to do some research before cleaning the data and choosing the variables.

Importance

The goal of this report is to understand the pattern of people's vote in the Canadian federal election. We aim to predict the next popular vote of the Canadian federal election with a regression model with post-stratification. This analysis could be important because the results could change the entire society and all residents. If the results from the analysis are significant enough, political parties can plan strategies to win votes. Other stakeholders like authorities and businesses can also take advantage of the analysis to cope with changes if one particular political party is likely to have more votes.

Background

There are many factors affecting one's decision to vote for a particular political party. Usually they include gender, race, region of residence, etc (Ma, 2006). Each of the factors might influence people's attitudes towards politics because of the values and culture. There are many factors affecting one's vote, and we hope to choose the most impacting factors for our regression model.

We processed the census data and survey data by matching the questions in the phone survey and the census data variables. We have studied the survey questions and decided the important variables that could be the main "game-changer" in the Canadian federal election.

Terminology

One of the political terminology we might introduce later would be partisanship, partisanship in politics represents one person's quality or action to support a political party without questioning or considering the entire circumstance or issue. There are also research showing that the increase of partisanship might shape the voting behaviour of people.

Research Question

The research question we are interested in is, "Do factors like region of residence, age, household size, education, household income, gender have an impact on one's vote for one particular political party?"

In the following section, there are descriptions of how we have studied and cleaned the data while providing some data statistics of selected variables. After that, we are presenting the methodology of our model and the assumptions we made. At last, results from our model will suggest whether the factors are significant enough to make an impact on vote's decision for one particular party.

Data

For our data, we have two separate datasets: one General Social survey (GSS) census data retrieved from the CHASS website and one survey data collected from the Canadian election study (CES). The census data includes responses to 81 questions that were asked in the 2016 census survey and the survey data includes responses to the 2019 CES phone survey conducted pre-election with over 278 variables.

To clean our data, we decided to only include six prediction variables in our dataset – thus, we kept only province, age, household size, education, family income, and sex. With these six variables, we have performed mapping, so that our response in each observation from the survey data matched with the census data. In our survey data, each variable represented a question from the survey with the labels "q1, q1, q3,. . . ", so we went into the documents from the CES website to locate what the questions represent. After that, we narrowed down the questions we needed for our model and subset them into a dataset. We also changed the headings to be more representative of each variable and match those in the census data. For some observations, the responses of each question have a specific number corresponding to a response. For example, in the variable "sex", it states "1" for male, "2" for female, "3" for other, "-8" for refused, and "-9" for "don't know". For questions answered in this way, we renamed the numerical variables with the phrase that it represents. We also grouped both education and income variables, so that the responses in our survey data match the census ones. In terms of age, the 2019 survey asked "what year are you born?" Therefore, we used 2019 and deduct the responses to calculate the age and stored those into our dataset. We also included four variables which are columns for each political party. Each row states "1" for their chosen political party and "0" for the other three they did not vote for.

To create our model, we chose to include one response variable (the chosen political party) and six predictor variables. The important variables of our data include the following: Province: what province the respondents currently live in Age: the age of the respondents Hh_size: household size Education: the highest level of education the respondents have completed Income_family: family income Sex: what sex respondents identified themselves as

Household Size

	median	mean	min	max
census	2	2.3467625	1	6
survey	2	2.6308125	1	15

From our observation, most people in both the census and the survey have a household size of two on average. We expect the survey to represent the census, so the number should be roughly similar. The mean should also be similar. However, the census has a cap on maximum amount of members in a household, that is not matched with the survey. Nevertheless, the overall amount of people exceeding that size 6 is 1.2774655 %. Thus, this does not make a huge impact on our prediction, but it should be aware as a source of error.

Age

	median	mean	min	max
census	54	52.182458	15	80
survey	51	50.8804292	18	100

Similarly, we see the median age is also ~52 years old, which is similar to the data. The mean is also quite close and similar to the data. There is a hard cap on census age of 80 years old as well, but the amount of people over that age 80 in the survey is 2.7337762 %. This tells us that it would not make a significant impact, but this is still a source of error. We have made an assumption that people of age 80 and 80+ all have relatively similar beliefs. As a result, we can consider this source of error not as impactful as the one above. On another note, we also see how Canada is an aging population with the median age being 52 that is close to the retirement age.

Sex

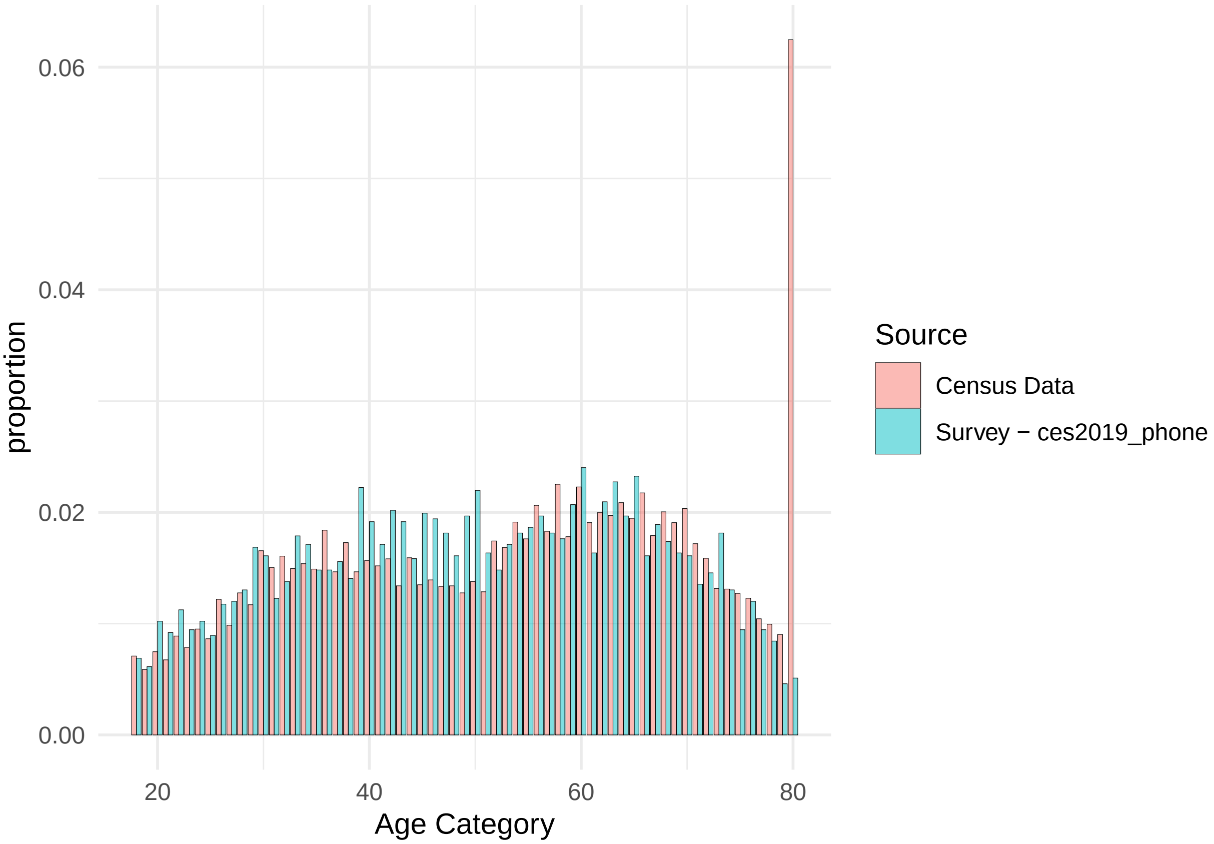
	prop male	prop female
census	0.4562178	0.5437822
survey	0.5625958	0.4374042

Next, we see the proportion of males is roughly equal and the compliment of female is also approximately the same. This is useful as we need the survey to roughly represent the population of the census. Therefore, we would like all summary statistics to be roughly the same.

Province

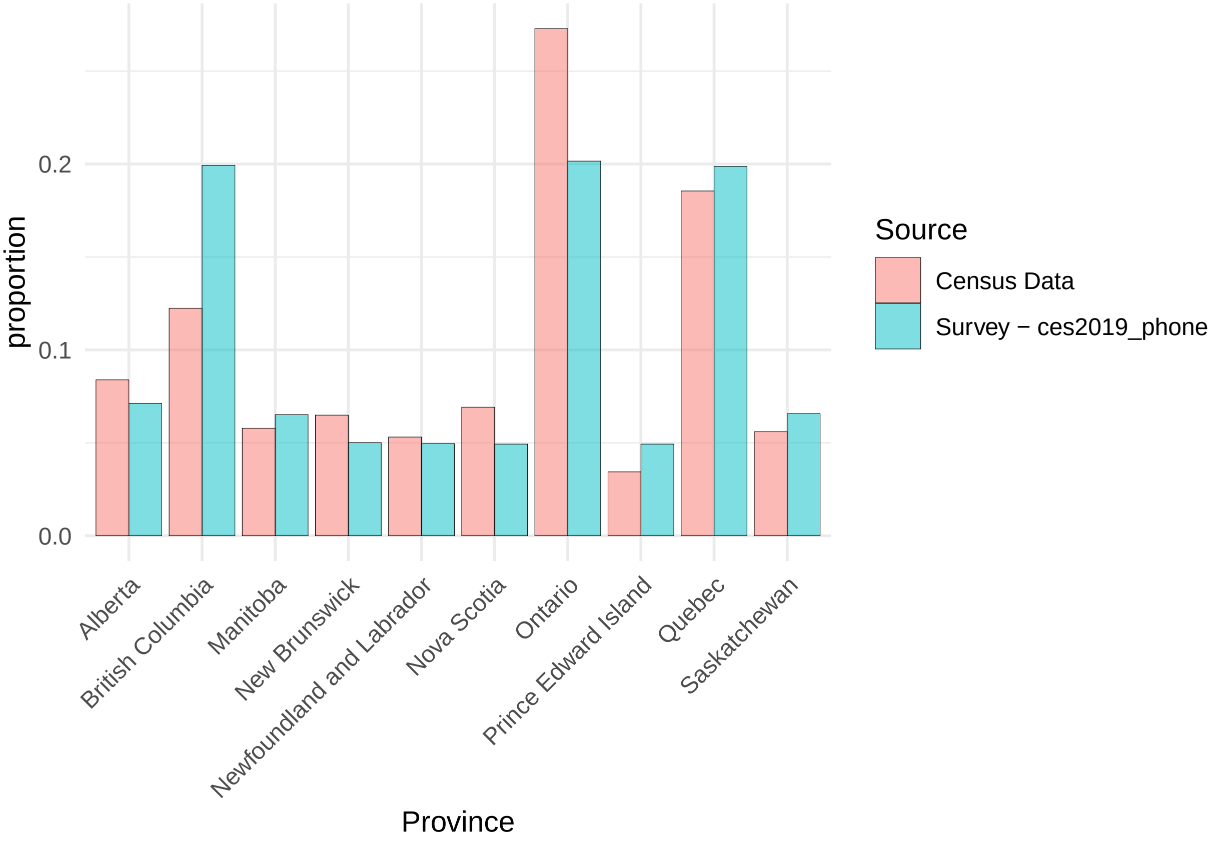
province	survey	census	ratio
Alberta	279	1728	6.193548
British Columbia	780	2522	3.233333
Manitoba	255	1192	4.674510
New Brunswick	196	1337	6.821429
Newfoundland and Labrador	194	1094	5.639175
Nova Scotia	193	1425	7.383420
Ontario	789	5621	7.124208
Prince Edward Island	193	708	3.668394
Quebec	778	3822	4.912596
Saskatchewan	257	1153	4.486381

The census size is 5.2636689 times bigger. If the ratio between census and survey is greater than 5.2636689, it tells us the survey is under representative. If it is lower, the survey is over representative. However, there are some larger deviations. Moreover, the same size of these provinces are low which results in more variance. In general, there is not any major problems, such as missing an entire province.

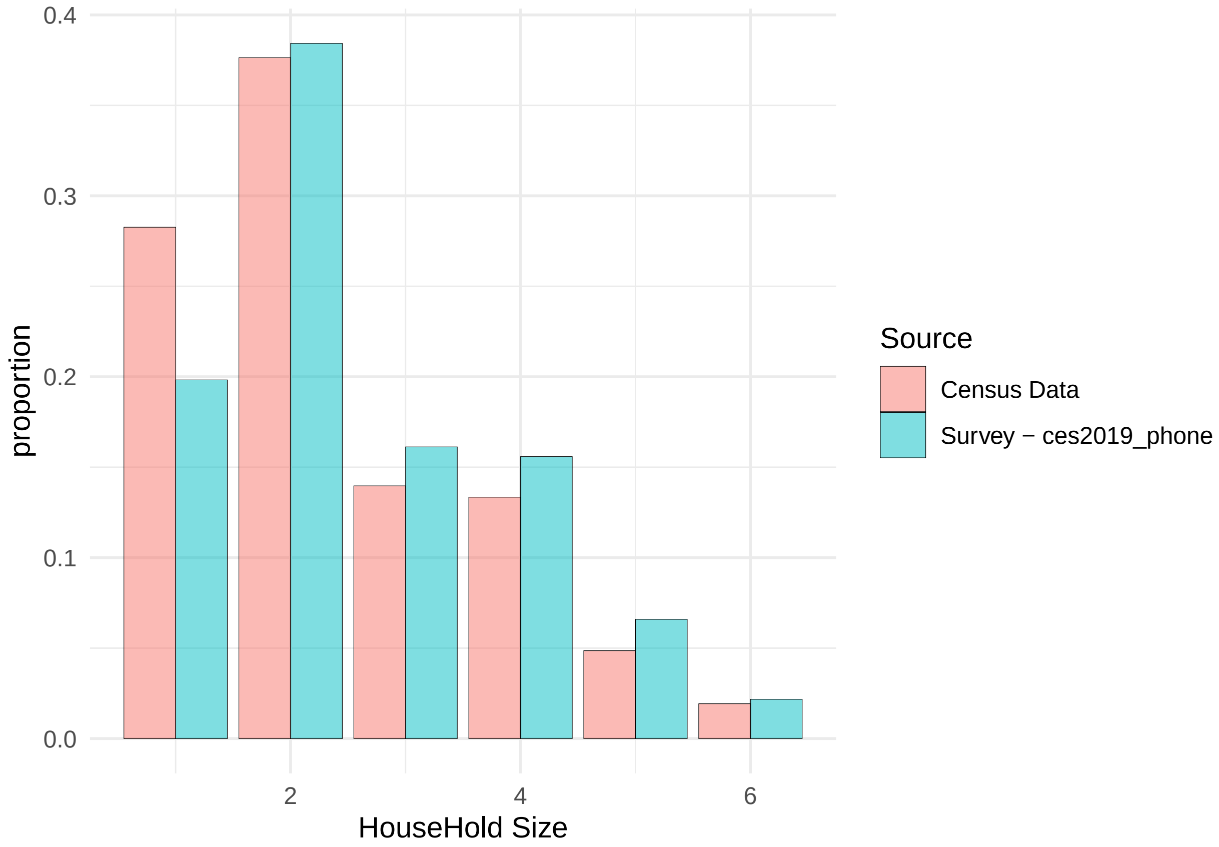


age bar graph

This graph shows us that both the age demographic is very similar between the census and survey. Also, there is not any exceptional spike to tell us otherwise except at the 80 year old region. As mentioned before, we believe that at the age demographic past 80, the result would not differ much. In other words, we expect a person of 80 years and a person of 95 to have a roughly same voting behaviour. This is an assumption that the elderly tend to support policies with support to people in the older years like retirement funds and care home policies.

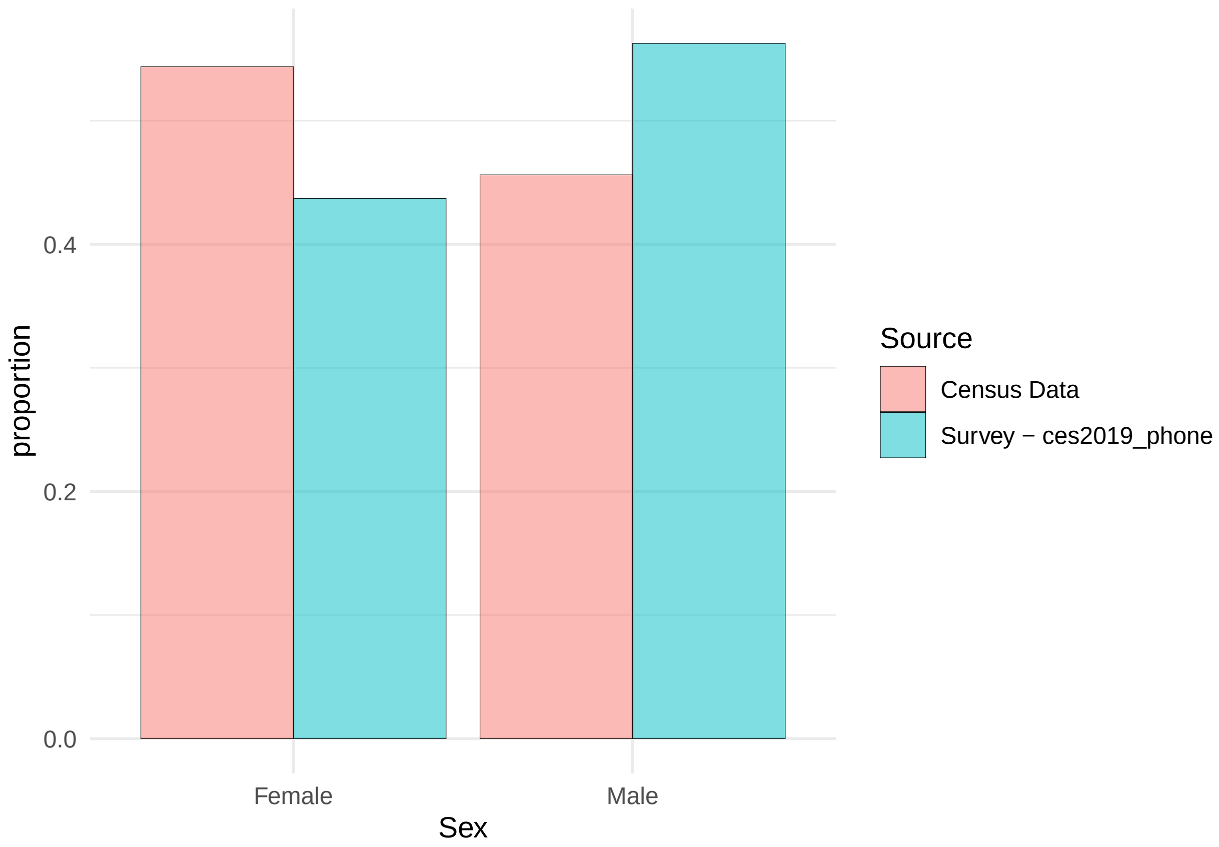


Province distribution bar graph Within the Province graph, we see that the proportions between the census and survey data are similar to each other, allowing a significant inference to be made. The respondents from Ontario, BC, and Quebec are also larger compared to other provinces. This could have a demographic influence on the political parties and their success rate from this survey.



Household size bar graph

In this graph, we can see that the proportions of household sizes are relatively similar between both the census and survey data. There is also evidence that there are less responses as the household size grows. This could be a good indication that larger families' preferences are different compared to those with a smaller household size.



Sex bar graph For the sex variable, both male and female have similar proportions. There are more female respondents in the census data while there are more male respondents in the survey data. Although this is the case, the difference is compensated relatively equally between both the survey and census data. Note that in this survey, only sex (Male and Female) was used.

Methodologies

The goal of our analysis is to predict the most popular election result for the next election in 2025. To achieve this goal, we have chosen several demographic characteristics and used them as variables in our regression model.

Assumptions

Based on the survey conducted in 2019, we have picked the four main political parties as our dependent variables. Leger had conducted a poll about voting intention in Canada, Liberal, Conservative, NDP are the three parties with the most voting intention (Renfrew, 2022). We have picked the third runner-up, Bloc Québécois as the dependent for our fourth model. Using the demographics characteristics might not be the most accurate factors to predict voting result, but the correlation between the demographics and election result had been shown to be stable over time (Kim, Zilinsky, 2022). There could also be a correlation between the demographics and the partisanship. When the partisanship has increased, the group division will increase as to the voting behaviour. Research also showed that there is a smaller correlation between the election forecast and the demographics data with social media (Sanders, et al., 2016). Therefore, we have used the phone survey result from 2019.

Moreover, we decided to select the most popular political parties, so that it can help us reduce the strongly influential outliers and variables, such as the less popular political parties. We also implement a logistic regression model for each of the parties instead of combining four of them together. This helps focus on the analysis for each party. We will understand if the demographic data have an impact on predicting the future election result. We have built separate logistic regression models, so that we can assume the response variable is binary. In this model, the sample size is sufficiently large with more than 3000 observations in the original dataset. We also assume that there is no multicollinearity among the explanatory variables. The age, household size, and province data do not have a highly correlation with each other. The observations of our analysis are independent of each other, there are random patterns across our variables.

These four models were selected in such a fashion that for each political party, we have a different model. Each model was selected from the numerous models made for the political parties. For each political party, one model was selected by comparing the p-values of the features, the F-statistic, and the adjusted R^2. Lastly, likelihood ratio test was conducted to compare the full and reduced model where it was necessary. This brought us done to the best model for each of the four political parties.

Parameter of Interest

We are interested in looking at the slope intercept of the logistic regression model as it has shown us a pattern where the voting behavior of each demographic group is. From our analysis, the slope value of our age intercept is more likely to be correlated to the voting party, which is our dependent variable. For example, if the observation is older, there is a higher probability that the voting result is Conservative Party.

Model Specifics

Describing the model Our model of choice is the logistic regression model. This model will determine the relationship between our variables of interest and the probability of choosing each political party. Since we are studying four political parties, we will create four separate logistic regression models.

The general model of a logistic regression is as follows:

$$log(p/1-p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

Where p is the probability of the event of interest (voting for the party of choice) occurring and the β coefficients represent the change in log odds for every one unit increase in the corresponding x (predictor) variable.

Conservative Model

The Conservative party model is formatted as

$$log(p/(1-p)) = \beta_0 + \beta_1 x_{ij} + v1_{oj} + v2_{oj}$$

where p represents the probability of the Conservative party being most voted for and β_0 represents the intercept. In $\beta_1 x_{ij}$, x represents the age and $v1$ is the random intercept which is household size while $v2$ represents the province.

NDP Model

For our NDP party model, it follows the format of

$$log(p/(1-p)) = \beta_0 + \beta_1 x_{ij} + v_{oj}$$

where p represents the probability of the NDP party being most voted for, β_0 represents the intercept, and in $\beta_1 x_{ij}$, x represents the age and v is the random intercept which is household size.

Bloq Quebecois Model

In our Bloq Quebecois Model,

$$log(p/(1-p)) = \beta_0 + \beta_1 x_{ij} + v1_{oj} + v2_{oj}$$

p represents the probability of the Bloq Quebecois party being most voted for and β_0 represents the intercept. In $\beta_1 x_{ij}$, x represents the age and $v1$ is the random intercept which is household size while $v2$ represents province.

Liberal Model

The Liberal party's model follows the format of

$$log(p/(1-p)) = \beta_0 + \beta_1 x_{ij} + v_{oj}$$

where p represents the probability of the Liberal party being most voted for, β_0 represents the intercept, in $\beta_1 x_{ij}$ x represents the age and v is the random intercept which represents province.

Post-Stratification

There are two ways we can predict the explanatory variable $\hat{y}^{PS}$. But, both ways start with running the observed values which are age, household size and province through the multiple logistic models.

Deterministic Approach

With the deterministic approach, we predict exactly what the value is and do not take the variation into account. We perform this by simply taking the highest probability predicted for each group and using it as our predicted value $\hat{y}^{PS}$. Supposing the parties probabilities are (0.75, 0.5, 0.1, 0.04) for the Liberal, Conservative, NDP, and Bloc Québécois, we will pick the Liberal party as the predicted value.

Probabilistic Approach

The other method is a more probabilistic approach, in which the model calculates all the probabilities of voting for each party. Assuming probabilities of voting for the parties are (0.75, 0.5, 0.1, 0.04) for the Liberal, Conservative, NDP, and Bloc Québécois respectively. At first, the list of probabilities did not add up to 1, so we have normalized it by adding the probabilities together and dividing each probability by it. Then, the total of the probabilities is in a total of 1. in our example, (0.75, 0.5, 0.1, 0.04) will be recalculated to (0.539, 0.36, 0.072, 0.029). After that, we will proceed to sample this value with the associated probability. We will then have a 53.9% chance where the person pick will be Liberal, 36% Conservative, 7.2% NDP and 2.9% Bloc Québécois. Then using these values we will aggregate the counts and determine the vote counts and then determine the percentage.

However, in both cases, the output will look something similar to the following.

party	count	prob
liberal	4	0.4
ndp	1	0.1
conservative	3	0.3
bloq quebecois	1	0.1

It is important to note that the deterministic approach would be undesirable due to the fact that we remove randomness and variance. For example, in the case of a coin that has 55% chance of heads and 45% chance of tails. Our model would predict that it has a 100% chance of heads, because it is always taking the larger value. Nevertheless, if we use the probabilistic approach to determine $\hat{y}^{PS}$, the result will show roughly 55% of heads and 45% of tails. This would create a more accurate prediction. Furthermore, we are going to show both methods in the result section below, and focus on the probabilistic approach as our “real method”.

Results

Probabilistic Approach

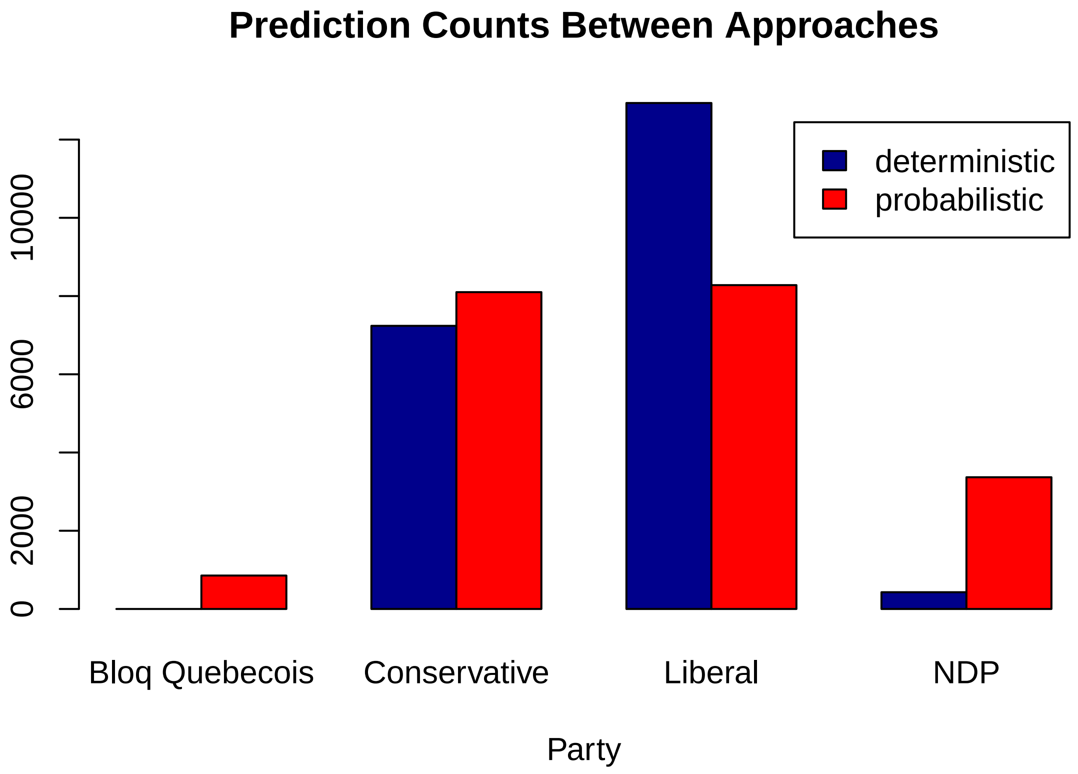
Party	count	percentage
Bloq Quebecois	855	0.0415008
Conservative	8100	0.3931657
Liberal	8279	0.4018542
NDP	3368	0.1634793

As mentioned above, this is our “main approach”. In our model, we only predicted 4 parties and skipped the remaining parties since the other parties only make up a minority in the total votes received. In this model, we are using the probabilistic approach as mentioned in the method section. We see that the Liberal party will receive roughly 40% of the votes, Conservative will receive roughly 39%, Bloc Québécois will receive 4% and NDP will receive 16%. This implies that the Liberals have the highest chance to win and the Conservatives will be close behind. Comparing to the 2019 election results, we see that the liberals had 40% of the vote, which is remarkably close to what we have predicted. On the other hand, the results for conservatives was lower than what we have predicted. Nevertheless, NDP and Bloc Québécois’s results were also close to our predictions. We recognize that our analysis did not include the made up for 3.45% of the votes. Therefore, most of the figures in our prediction could be of a slight over estimation. However, due to Canada’s system of having seats in parliament, the popular vote or party who was voted the most may not win the election still.

Deterministic Approach

Party	count	percentage
Conservative	7238	0.3513251
Liberal	12935	0.6278517
NDP	429	0.0208232
Bloq Quebecois	0	0.0000000

After the probabilistic approach, we are going to look at the deterministic approach. This model performed terribly compared to the actual results in 2019. The model showed that the liberals would have absolutely won the election. In the last section, although the model predicted that the individual would vote for liberal, the probability of voting conservative eventually might only be a slight less. As a result, parties may not get the full representation they deserve.



Finally, looking at the red bars in the graph, the Liberal and Conservative parties are the only parties that will have a chance at winning the election. As mentioned in the deterministic approach, Bloc Québécois naturally has a close to 0% chance of being voted due to a much lower probability when compared to the liberal party.

Conclusions

In our report, we performed four separate logistic regressions based on the phone survey data pertaining to our six variables of interest: Province, Age, Household Size, Education, Family Income and Sex. Through post-stratification, we were able to use our logistic regression's results from the survey data and compare it to the larger census data. The results of the post stratification indicate that the Liberal party will have the most votes in the upcoming 2025 election, although the Conservative party is coming in a close second – both with nearly 40% of votes in their favour.

Limitations

Our first limitation of the analysis includes the income portion. In the CES survey data, the question asks about total household income which could be different with the question asked in the GSS census data. This could cause some ambiguity to understand the truth behind the responses to this question. Moreover, these surveys were conducted during two different time frames, one in 2016 and the other in 2019. Due to socioeconomic differences between these two time frames, there could be some unrepresentative numbers in terms of comparing incomes.

One limitation in our analysis was the difference between the categorization of the variables between datasets. In the survey data taken from the CES survey, respondents were asked to give their "gender" while the GSS census data asked for "sex". Ultimately, these two terms are used interchangeably while they should not be. Sex pertains to the more physical biological aspects while gender refers to more of the identity and behavioural characteristics. For our analysis, we address the mapping of these variables by including only those who responded either "male" or "female" within our model.

Finally, we decided to eliminate the Green and People's Party of Canada parties due to their unpopularity in the past elections. Since we are just trying to predict the next elected political party in Canada, we decided not to reference them in this report, although their extrapolation would follow the same method and model as the others.

Next Steps

In the future, any analyses exploring the relationship between province, age, household size, education, family income and sex should be determined using data that is 'mappable' to one another. The questions should also be the same without any ambiguity as to how the readers may interpret the questions. From our analysis, we found that the two major runners for the new political party are those of the Liberal and Conservative parties. These make sense and are similar to previously hypothesized numbers and follow the trend of expected election results.

Another thing to keep in mind is that our model uses four separate logistic regressions for each political party. Another way to do this would be using a multiple logistic regression. This way we would have one large model instead of four separate models. In doing so, there would be less opportunities for errors, violations, etc. while also keeping similar assumptions.

Final Remarks

The Canadian Elections are a chance for people to exercise their rights to vote. Through this analysis, we understand more of people's behavior and understanding of who may prefer one political party over another. People will also choose to vote for political parties that have attractive plans for policies and the results of each election have a detrimental effect on Canadians and their economy. From these experiments, we can begin to understand and predict confidence in things, such as the election, that have such a large impact on the nation and its people.

Bibliography

1. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)

2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

3. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)

4. Kim, S. S., Zilinsky J. (2022, August 20) . Division Does Not Imply Predictability: Demographics Continue to Reveal Little About Voting and Partisanship. Springer. https://link.springer.com/article/10.1007/s11109-022-09816-z.

5. Ma, C. (2006, February 7). *Voting behaviour in Canada.* The Canadian Encyclopedia. https://www.thecanadianencyclopedia.ca/en/article/electoral-behaviour. (Last Accessed: November 20, 2022)

6. N.d. (2021, April 29). *2020 presidential election voting and registration tables now available.* The US Census Bureau. https://www.census.gov/newsroom/press-releases/2021/2020-presidential-election-voting-and-registration-tables-now-available.html. (Last Accessed: November 20, 2022)

7. Renfrew, M. (2022, November 15). Liberal 32%, Conservative 34%, NDP 19%. LÉGER. [https://cultmtl.com/2022/11/liberal-32-conservative-34-ndp-19-leger-poll-federal-election-voting-intentions-canada/] (https://cultmtl.com/2022/11/liberal-32-conservative-34-ndp-19-leger-poll-federal-election-voting-intentions-canada/).

8. Sanders, E., et al. (2016, November) Using Demographics in Predicting Election Results with Twitter. ResearchGate. https://www.researchgate.net/publication/309272639_Using_Demographics_in_Predicting_Election_Results_with_Twitter.