Predicting Popular Vote for the upcoming Canadian Federal elections of 2025

STA304 - Assignment 3

Group 47: Jay Kanchankoti, Stuti Sekhri, Jiyun Yoo, Janhavi Agarwal

November 5, 2021

Introduction

In recent years, forecasting election has been drawing attention from academia and public to predict the outcome of the election that impacts the society in the future. While there exists several methods of prediction from political perspective to data-driven approach, it is important to utilize accurate method to analyze the election data in a unbiased environment. Since it is crucial to have reasonable standards to use in determining the advantage of various models, here we explore the General Social Survey (GSS) as the "census" data (Technology, Computing in the Humanities and Social Sciences 2019), and data from the CES2019 package as "survey" data (Government of Canada, 2017). Given the data resources available from Canadian Election Study, Phone Survey data for the year 2019, it is reasonable to hypothesize that Conservative party would win based on the overall popular vote of the previous Canadian federal election.

Data

The data used for the purpose of this report is the Canadian Election Study, Phone Survey data for the year 2019.This data is referred to as the survey data in the report.The reason that the data from 2019 was taken was due to the fact that it was the most recent survey data available and we wanted the factors affecting the popular vote to be the most relevant. This CES data was obtained from the 'CesR' package already installed int the Rstudio. The CES has been a pivotal source of data on Canadians' political behavior and attitudes, measuring preferences on key political issues. The data touched on issues like the income of the respondent, their social behavior towards the different factors in the society, their political inclinations and feelings etc. This data provide an unparalleled snapshot and record of Canadian society and political life. Another data relevant to the report is the General Social Survey data. This data was obtained from http://www.chass.utoronto.ca/ which is the website for Computing in Humanities and Social Sciences.We are using the 2013 version of this data because we didn't want to factor in the changes in the vote due to COVID 19 pandemic. The GSS data is referred to as the census data in this report and it is significant because it monitors the living conditions and the social well being of Canadians. This data helps in making policies as it provides a comprehensive look at a variety of essential topics like care giving, families, time use, social identity, volunteering and victimization.

Data Cleaning Process To clean the data we selected only the variables in both the datasets which were common and significant to our analysis. Since the objective of this report is to predict the popular vote of the next Canadian Elections with a regression model, the variable we chose were age, gender, income, education, province and the party they voted for. The purpose of selecting the variables which are common in both the datasets is because later we are going perform a Post-Stratification and for that we need the variables which we can map from the census data to the survey data. For cleaning the income variable we categorized the data into groups of range of income and similarly for the education variable we grouped them

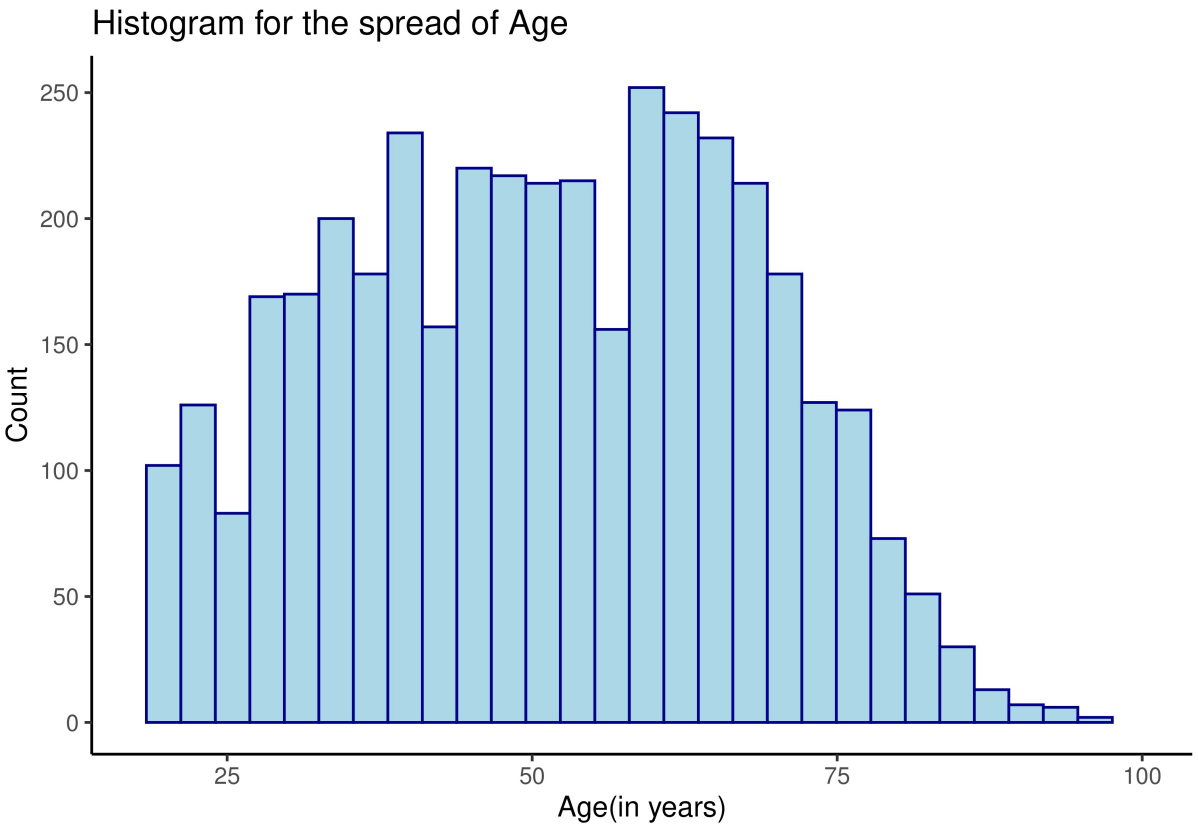
into categories to make them easier to work with. This whole cleaning process was replicated in both the datasets for convenience of Post-Stratification.

The important variables for the analysis after the p value test are Age, Income and Province.

Numerical Summary of Age (Quantile)

0%	25%	50%	75%	100%
15.000	37.300	54.200	66.775	80.000

The above table indicates the quantile range of the variable age, it shows that the median age for the age is 54.2 the spread of the age is from 15.0 (but since the legal age to be eligible to vote is 18 in Canada for the model later in the report we are going to drop the age range before 18) to 80.



The above histogram shows us the spread of age, we observe that the histogram is undefined and multi-modal. the histogram being multi-modal refers to the fact that there is more than one mode or score that occurs most frequently. This also helps us understand that most of the people surveyed are belonging to the age group 37 - 70 as most of the histogram seems to be clustered there.

Methods

The goal of this study is to predict the popular vote based on general characteristics of the population such as age, gender, income and so on. A logistic regression model shall be created to predict the probability of vote for a particular party by a group of people with certain characteristics. A logistic model because the output, vote for a particular party, is binary – yes or no. This model will be created based on the

data available through the phone data of the Canadian Election Study, 2019. This model will be run thrice to check the probability of vote for the Liberal, Conservative and the NDP party. These three parties are chosen specifically as historically, these are the parties with the most popular votes (Hahn, 2021). These probabilities would then be mapped on to the same groups of people in the census data to get the predicted percentage of vote for each of the three parties.

Model Specifics

Below is the summary statistics for all the variables used to model the probability for a liberal vote. The model will be selected based on whether the variables have a p-value less than 0.05.

We can see that the only variables with a p-value less than 0.05 are the intercept, age, education and province. One of the assumptions in this study is that the model created is using only the votes for the Liberal party as the output and that the same model is accurate for predicting the probability of votes for other parties. Therefore, our new model will look as follows for each of the three parties:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1(Ages31-45) + \beta_2(Ages46-60) + \beta_3(Ages61-80) + \beta_4(education) + \beta_5(BritishColumbia) + \beta_6(Manitoba) + \beta_7(NewBrunswick) + \beta_8(Newfoundland) + \beta_9(NovaScotia) + \beta_10(Ontario) + \beta_11(PEI) + \beta_12(Saskatchewan) + \beta_13(Quebec)$$

Where y represents the the probability of vote for the party, β_0 represents the intercept, $\beta_1, .., 13$ are the co-efficients for the respective x variables where the x variables are age groups, education level and province.

Post-Stratification

Now that our models have been established, we can use these probabilities, of voting for a certain party based on the characteristics of a group of people, and multiply it with the proportion of the population in that group. Summing up these weighted probabilities on the different groups would give us the popular vote for each party. This process of splitting up the population into groups and multiplying their probabilities from a sample data, in this case the survey data, is called the post-stratification method. These groups are often called "cells" in statistics. The formula is given below:

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

Where $\widehat{y}_j$ the estimate in each cell and $\sum N_j$ is the population size of the jth cell based off demographics. The cells in this study are grouped off in age groups of 18-30, 31-45, 46-60, 61-80, in education of less than and more than equal to Bachelor's Degree and by each province. These groups were bifurcated in such a way because they give the most significant results for the voting probability.

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

Results

Model for Liberal Party On running our logistic model for the Liberal party, this is the equation we obtained:

$$log(\frac{p}{1-p}) = -2.689 + 0.0243(Ages31-45) - 0.0535(Ages46-60) + 0.2958(Ages61-80) + 0.5517(education) + 0.8356(BritishColumbia) + 0.9283(Manitoba) + 1.037(NewBrunswick) + 1.39(Newfoundland) + 1.301(NovaScotia) + 1.471(Ontario) + 1.344(PrinceEdwardIsland) + 0.1909(Saskatchewan) + 1.138(Quebec)$$

Model for Conservative Party On running our logistic model for the Liberal party, this is the equation we obtained:

$$log(\frac{p}{1-p}) = 0.06515 + 0.17962(Ages31-45) + 0.50277(Ages46-60) + 0.47531(Ages61-80) - 0.44329(education)$$
$$-1.40829(BritishColumbia) - 0.6539(Manitoba) - 1.28255(NewBrunswick) - 1.86698(Newfoundland)$$
$$-1.53218(NovaScotia) - 1.38154(Ontario) - 1.58506(PrinceEdwardIsland) - 0.40255(Saskatchewan) - 2.27396(Quebec)$$

Model for NDP Party On running our logistic model for the Liberal party, this is the equation we obtained:

$$log(\frac{p}{1-p}) = -1.1779 - 0.79536(Ages31-45) - 1.20492(Ages46-60) - 1.57529(Ages61-80) + 0.07847(education)$$
$$+0.95542(BritishColumbia) + 0.47405(Manitoba) - 1.09471(NewBrunswick) + 1.03097(Newfoundland)$$
$$+0.59163(NovaScotia) + 0.64228(Ontario) - 0.64934(PrinceEdwardIsland) + 0.48271(Saskatchewan) - 0.11902(Quebec)$$

Given below is the table with the post-stratified values of the $\hat{y}^{PS}$ for each of the three models.

$\hat{yPS}_{liberals}$	$\hat{yPS}_{conservatives}$	$\hat{yPS}_{NDP}$
0.247460	0.291490	0.059361

From the table above we see that the conservative party has the highest number of popular votes at 29.15% based on the variables we selected. This is followed by the liberals at 24.75% and NDP at 5.94%. These results seem reasonable based on last elections popular vote where a similar ranking followed.

Conclusions

As forecasting elections research has been a significant factor of predicting the potential in our society, the research in academia and attention in public has been impacting various phenomenon. Based on the overall popular vote of the Canadian federal election data, we hypothesized that the Conservatives would win the next federal election.

In order to predict the probability of vote for a certain party by a group of people, we created a logistic regression model since the output from the voting results is binary. The model run thrice to see the probability of vote for three parties, the Liberal, Conservative and the NDP party, chosen by the most popular votes (Hahn, 2021). Thus, the model allowed us to predict the voting percentage of the vote from the each party.

After we established the model, the probabilities of voting for certain party based on different demographics of the group of people such as age and education. We were able to obtain the popular vote by multiplying with the proportion of the population of the group and summing up the weighted probabilities on the different groups. The post-stratification method was utilized in this process since we spitted up the population and multiplying the probabilities of the group in the survey data.

The main results of this report were that we found out after the regression model and post stratification is that the conservatives have the popular Canadian Vote with 29% of the proportion of Canadians followed by Liberals with 26% of the proportion of Canadians and lastly the NDP party has around 9% of the proportion of the popular vote. These results were calculated by looking at the variables age, income and province.

This report focuses on the big picture of the Canadian political scenario in the year 2025. But if there is an early election recalled this report is still valid.

Where this report falls short is on its assumptions. Since we only took two variables significant to predict, namely age and education; these are not enough to ideally predict the popular vote for the next federal

elections. The voters decide on who to vote based on a lot of things, like the political ideologies of the party, the work they have done in the past etc or the decision could be completely random. This report fails to take in all these considerations to make the prediction.

The next steps recommended for the future reports on the census data for the prediction of popular vote is to consider and incorporate more and better variables that affect the voters decision to vote and get a better more varied population for the census so that all of the ideologies are covered.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (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. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)

4. Government of Canada, S. C. (2017, February 27). The General Social Survey: An overview. Government of Canada, Statistics Canada. Retrieved November 6, 2021, from https://can01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww150.statcan.gc.ca%2Fn1%2Fpub%2F89f0115x%2F89f0115x2013001-eng.htm&data=04%7C01%7Cjay.kanchankoti%40mail.utoronto.ca%7Cbcab7eef082643fa3a4708d9a0e984b9%7C78aac2262f034b4d9037b46d56c55210%7C0%7C0%7C637717747467362674%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1
3D%7C1000&sdata=5pm7HFwstzyf1tkPVCYKKe%2BqhHND6Go9pRiQo1%2F8t1M%3D&reserved=0.

5. Hahn, P. (2021, August 20). Interactive: How Canadians voted in the past 7 federal elections. CTVNews. Retrieved November 6, 2021, from https://www.ctvnews.ca/politics/federal-election-2021/interactive-how-canadians-voted-in-the-past-7-federal-elections-1.5553874.

6. Technology, A. K. through. (n.d.). Computing in the Humanities and Social Sciences. Retrieved November 6, 2021, from http://www.chass.utoronto.ca/.