What Could Be Different If Everyone Had Voted in the 2019 Canada Federal Election? The Conservative Party Could Have Won!

Using logistic regression model with post-stratification to retrospectively predict the 2019 Canadian Federal Election outcome and explores the myth of the turnout rate.

Shang Wu 1004797428

December 22, 2020

Github Repository

Code and data supporting this analysis is available at: https://github.com/wsshirley/STA304-Final-Project

Abstract

In this report I focus on how the turnout eate could be an important factor in determining the outcome of a election. To predict what could be a different outcome in the 2019 Canadian Federal Election, I analyze the proportion of voters that voted for one of the six parties, assuming that everyone had voted. To complete this goal, I build six generalized logistic regression models for the six parties using a survey dataset and apply post-stratification techniques on a census dataset to calculate the population-level proportions. As a result, it turns out that we could get a very different election outcome. That is, the Conservative Party could have won the election if there was a high turnout rate for the election.

Keywords

Logistic regression, Post-stratification, Survey data, Census data, Canadian Federal Election, Parties, AIC, BIC, Receiver Operating Characteristic Curve, Area Under Curve.

Introduction

The Canada Federal Election can be considered one of the most important events that can affect Canadians' life and the operation of the government. Since the governing party represents the highest authority of Canada and it makes crucial decisions such as where the taxpayer's money is used, and how laws and policies are formulated, Canadian citizens cannot care less about the election results. However, the voter turnout remains around 77% as compared with 2015 election, and it has a decreasing trend as more Canadians became less interested in politics or more people report disabilities (Statistics Canada). Therefore the outcome only represents the preferences of people who actually voted, without accounting for the choices of the remaining quarter of populations. Hence it is of our interest to explore what could be different if everyone had voted in the 2019 Canadian Federal Election.

One of the most popular ways to make good corrections of a non-representative population is to use a regression with post-stratification (Stan User's Guide). It is carried out using both a survey data and a census data in a two-step process. In this project, I will first build a generalized logistic regression model assuming everyone had voted in the 2019 election, then apply post stratification techniques to calculate and weigh the estimates in each demographic cell to extrapolate how the 2019 Federal Election outcome could be different if everyone had voted.

Two datasets will be used to retrospectively predict the outcome in the 2019 Federal Election; one will be used to build the models and the other one will be used for post stratification. In the Methodology section (section 2), I describe the data obtaining and cleaning process, and the steps taken to build and validate the regression models for each party by selecting the most important predictors. Then I perform a post-stratification analysis to predict the outcome. Results of the analysis are provided in the Results section (section 3), and summaries of the data along with conclusions are presented in the Discussion section (section 4).

Methodology

Data

In order to make an analysis using a logistic regression model with post-stratification, a survey data and a census data are required. The survey data is obtained from the telephone survey in the 2019 Canadian Election Survey package (Paul A. Hodgetts and Rohan Alexander, 2020), and the 2017 Canadian General Social Survey from the CHASS data centre of University of Toronto is used as the census data. All of the cleaning and modeling processes are accomplished in R (6-12).

For the survey data, the target population is all Canadian citizens 18 years of age or older who reside in one of the 10 provinces; the sampling frame is the list of all residential telephone numbers in Canada, and by using the "dual-frame-with-overlap" approach for the landline frame and the cell frame, the researchers can incorporate weight to the overlap of these two frames to correct for higher and biased selection proportion if a voter is in both frames. The sampled population would just be anyone who responded phone calls and participated in the survey. People who are not-in-service and non-residential are automatically excluded. The Computer Assisted Telephone Interviewing (CATI) was used for phone surveys and a two-phase data collection was carried out, including a Campaign-Period SURvey (CPS) and a Post-Election-Survey (PES) (Canadian Election Study, 2019, Phone survey).

The data collecting process is quite similar for the census data: the target population is all non-institutionalized people who are 15 years of age or older and living in households in the 10 provinces of Canada; the sampling frame is the list of telephone numbers and dwellings available to Statistics Canada; and the sample population is all interviewees who answered questions through a phone call. The CATI method was also used here. As for non-responding interviewees, the researchers adjusted the weight of their responses to make up for the overall representativeness. This census data is collected by adopting the stratified random sampling on different areas, which includes conducting simple random sampling in each strata. This can reduce the bias of the sample; however, inefficiency and inappropriate strata divisions can make the sampling process problematic (2017 General Social Survey: Families Cycle 31).

The survey data has 4,021 observations with 278 variables in question form, while the census data has 20,602 observations with 81 variables. Having large data size can moderately reduce biases and increase accuracy of the analysis; nevertheless, too many categorical variables and variables with large missing proportions requires our attention. Since post-stratification can only be applied when these two datasets have common variables, only some of the relevant variables are selected and the pre-cleaned census dataset (Rohan Alexander and Sam Caetano, 2020) was used for further cleaning steps. In particular, the levels of categorical variables are standardized for both datasets so that all common variables have the same subgroups. Furthermore, the response variables are dichotomized by creating new binary variables indicating a voter's choice of party; for example, whether the voter voted for the Liberal Party or not. Details of cleaned variables are shown in Table 1.

Table 1: Variable Description Table

Variable Name	Description	Categories
age group	the age group the voter belongs to	[18-24], [25-34], [35-44], [45-54], [55+]
sex	the sex of the voter	male/female
birthplace	the birthplace of the voter	Canada/Other
province	the province the voter currently lives in	Alberta, British Columbia, New Brunswick Nova Scotia, Newfoundland and Labrador Manitoba, Ontario, Quebec Prince Edward Island, Saskatchewan
marital status	the marital status of the voter	married/not married
education	the education level of the voter	university or higher degree/below university degree
income family	the household income of the voter's family	less than 25,000, [25,000-49,999] [50,000-74,999], [75,000-99,999] [100,000-124,999], 125,000 and more

The variables that are common in both datasets describes some geographic and demographic characteristics of the observations. After removing missing values, I regrouped some categorical variables into fewer but more representative groups. For the case of having similar variables such as region and province, I only choose one of them for simplicity and to avoid multicollinearity. Also, to satisfy the assumption that "everyone had voted", and to reduce biases, I deleted observations that did not decide which party to support or refused to respond. Moreover, some continuous variables in one dataset are grouped according to the categories in another dataset for standardization; for example, age in the census dataset is converted into categorical variables with corresponding subgroups as in the survey data.

Details of some relatively important variables were plotted in Plot 1-4 in Figure 1. The coloured bar plots represent the vote count in different groups for each party (1-Liberal, 2-Conservative, 3-NDP, 4-Bloc Quebecois, 5-Green, 6-People). Generally speaking, it is obvious that the majority of votes are for the Liberal and the Conservative parties, with the People's party being the most unfavorable one. Looking in more details, almost half of the population who voted for the Liberals and the Conservatives are of age 55 and older (Plot 1). In addition, it is evident that more males voted for the Conservatives party than females (Plot 2); also, almost all the voters are born in Canada, especially for everyone who voted for the People's party (Plot 3). From Plot 5 of Figure 3 in the Appendix, it is notable that more people with a university or higher degree voted for the Liberal party, whereas more people without a university or higher degree voted for the Conservatives party. It is also an expected fact that all voters for the party Bloc Quebecois are from Quebec (Appendix Plot 7). More information of the variables can be extracted from the figures.

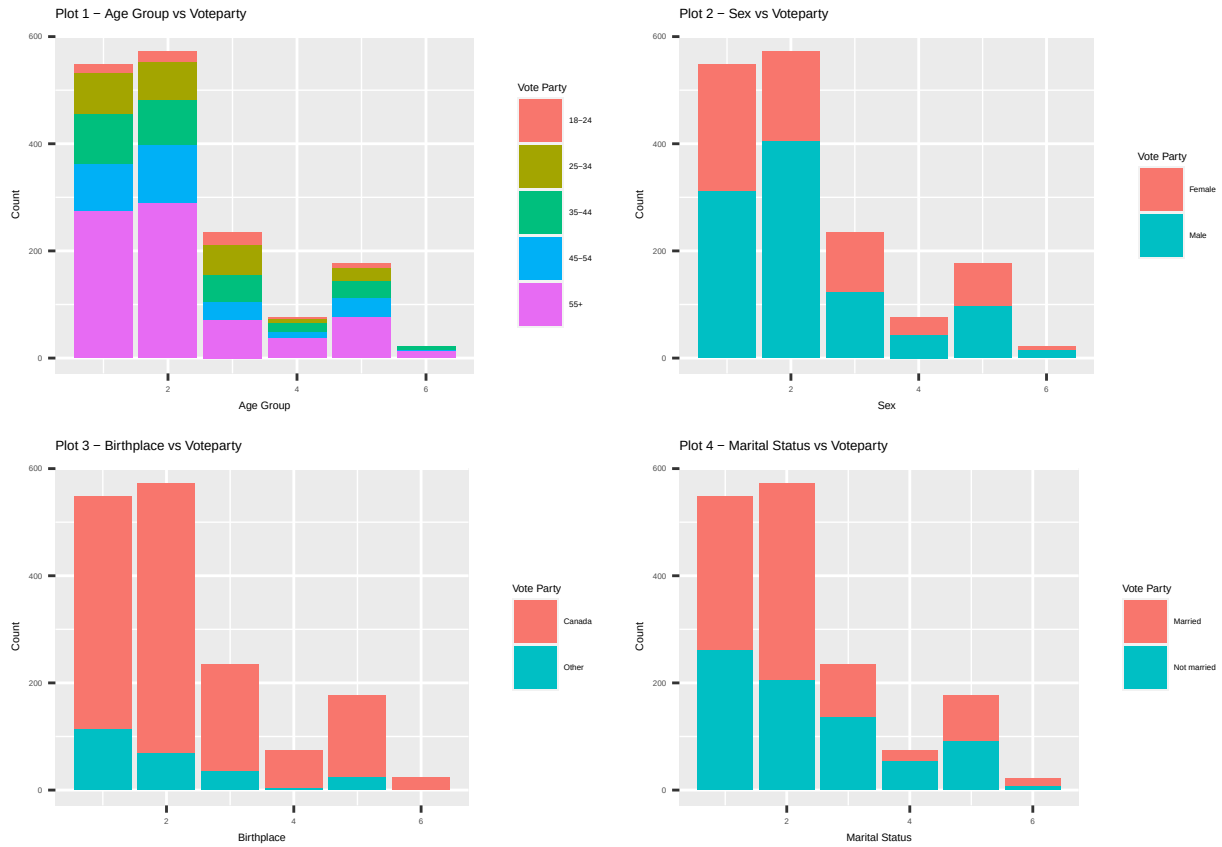

Figure 1: The barplots of four of the most representative variables

Model

The goal of this analysis is to see what could be a different outcome if everyone had voted in the 2019 Canadian Federal Election. Using the survey data, I built six logistic regression models for each candidate party and employed post-stratification techniques using the census data to retrospectively predict the proportion of votes for each party. In the following subsections, I will describe the process of obtaining the six final models as well as the post-stratification calculation.

Since I dichotomized the response variables, it was appropriate to use a generalized logistic regression to model relationship between the log odds of the probability that voters vote for a particular party in Canada. The general formula for this model is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... \tag{1}$$

Where p is the probability for voters to vote for a party, and the left hand side is the log odds of the probability. The right hand side consists of an intercept β_0, which represents the probability of voters would vote for the party defined in p when all other variables are zero (i.e. in their reference level). It also consists other predictors x_i with their corresponding coefficients β_i's; here β_i means when $x_i = 1$, we expect the log odds of p will change by β_i compared to the log odds of p when x is in its reference level, holding all other variabled fixed.

Among the common variables in both datasets, I selected seven most representative variables based on my preferences (see Table 1), including demographic characteristics such as age group, sex, birthplace, marital

status, education, family income, and geographic factors such as living province. According to the bar plots of these seven variables shown in Plot 1-7 of Figure 1 and Figure 3 in the Appendix, there are significant differences between the voters' opinions regarding different parties in each category; thus there is a high possibility that these factors would contribute to different final results. Therefore, these variables were used to create the full models for six different response variables, where each of the responses represents the log odds of the probability of voters vote for that particular party.

After the full models were created, I used AIC and BIC selections to further narrow down significant variables. Each of the six models will be reduced by stepwise selections, and the model produced by either AIC or BIC was kept as the final model for each party. Since AIC selection usually keeps more variables than BIC selection does, due to BIC's harsher penalty on the number of variables, I preferred to use models resulting from BIC selection for simplicity. However, if the prediction process was interfered because there were too few variables left after BIC selection, I would use AIC instead (for example, the final model for Green party and People's party were from AIC selection). Each of the six final models should have around 3 to 4 variables left, and the unselected models could be saved as alternative models. The alternative models were not selected mainly because they either had too much or too few predictors, which could make the models too complex or too simple that they were less accurate. But sometimes having more predictors can increase the precision of the predictions, despite the fact that more data has to be collected to use the model.

After obtaining the final models for all the parties, they were validated using the Receiver Operating Characteristic Curves (ROC Curve) with corresponding Area Under Curve (AUC) values. Since the x-axis represents the false positive rate (FPR) and the y-axis represents the true positive rate (TPR), it is expected that the curves are closer to the upper left corner, which means higher TPRs and lower FPRs. Correspondingly, higher TPR results in a larger area under the curve, thus larger AUC values (closer to 1) are expected as it indicates better model performance. As shown in Figure 2, five of the models had AUCs ranging from 0.64 to 0.75, while the remaining one had an unrealistically high value (0.93) and a sharp corner (ROC for Bloc Quebecois). This abnormality will be discussed in the weakness section. In general, the models' performances were acceptable but high accuracy and precision of predictions from these models might not be expected.

To retrospectively predict the result of the 2019 Canadian Federal Election, the last step was to perform a post stratification analysis on the census data. The census data was first partitioned into small cells (2,920 in total) based on demographic groups (all seven variables were used for partition), and then the proportion of voters who vote for a certain party was calculated for each cell. Next, these cell-level estimates of proportions were weighted by the size of cells and were aggregated to give a final estimate of the national-level proportions of voters that will vote for each of the six parties in Canada. The following formula was used for post-stratification estimates,

$$\hat{y}^{PS} = \frac{\sum_j N_j \hat{y}_j}{\sum_j N_j} \tag{2}$$

where $\hat{y}^{PS}$ represents the national proportion of votes for each candidate party; $\hat{y}_j$ is the estimated proportion of voters for a given party in the j^{th} cell, and N_j is the population size of the j^{th} cell in the census data.

Results

From the methodology section, I yielded six generalized logistic regression models, where each model predicts the log odds of the probability that voters vote for a particular party. The estimated coefficients of predictors in each model are shown in Table 2. Notice that the province variable appeared in four of the six models; sex, birthplace, marital status and education also showed up very frequently. In comparison, age group was only in the model for the NDP party, and family income did not show up in any of the models. This suggests that voters' age and family income are unlikely to affect the choice of party that they would vote for. To look more specifically at the final models and their interpretations, I will explain the model for the Liberal party as an example, and other models can be interpreted in a similar way.

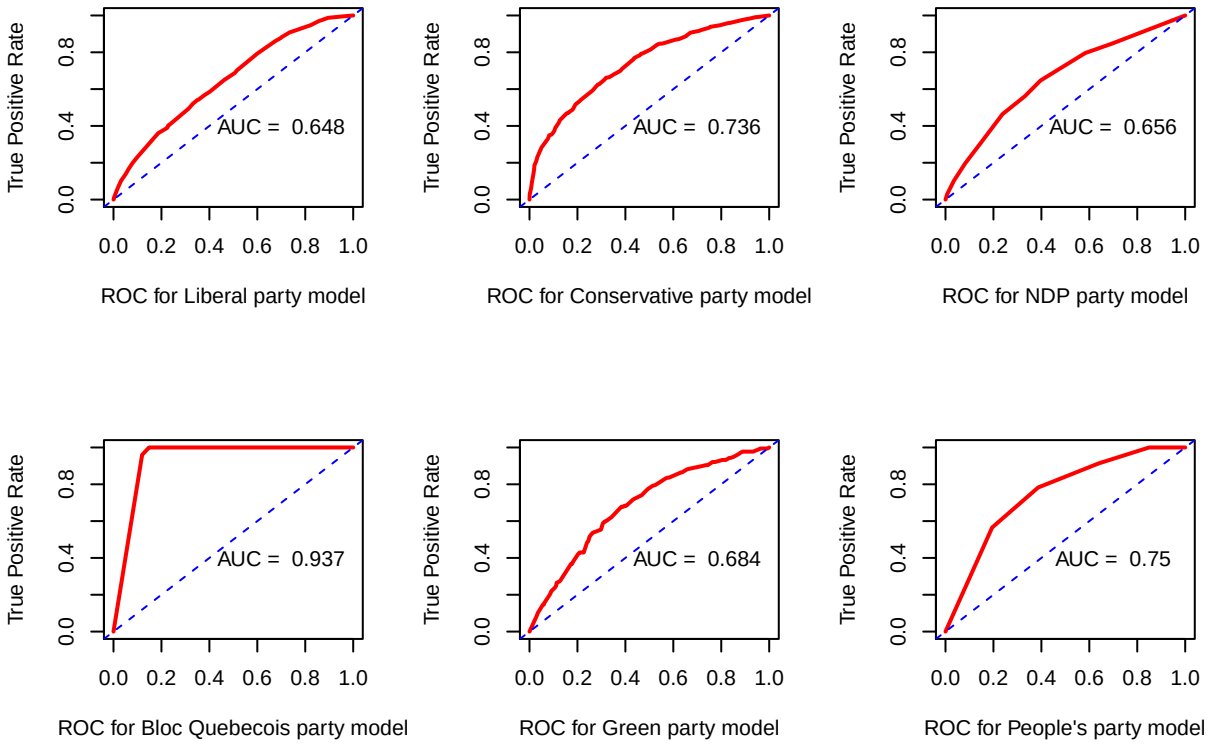

Figure 2: ROC Curves with AUC values for six models

The model for the Liberal party with estimated coefficients (obtained from first column of Table 2) is:

$$\begin{aligned} log(\frac{\hat{p}}{1-\hat{p}}) = &-2.212 + 0.613x_{other} + 0.833x_{BC} + 1.011x_{MB} \\ &+1.235x_{NB} + 1.669x_{NL} + 1.530x_{NS} + 1.523x_{ON} \\ &+1.578x_{PE} + 1.322x_{QC} + 0.232x_{SK} + 0.496x_{University} \end{aligned} \tag{3}$$

Where

- $\hat{p}$ is the estimated probability that a voter votes for the Liberal party;
- -2.212 is the intercept representing the log odds of p when a voter was born in Canada, lives in Alberta, and with education level below university degree;
- 0.613 is the coefficient for the dummy variable x_{other}. $x_{other} = 0$ is the reference level indicating a voter was born in Canada; $x_{other} = 1$ means a voter was born outside Canada and 0.613 represents the expected change in log odds of the probability that a voter who was born outside Canada votes for the Liberal party is 0.613 compared with the log odd for voters who was born in Canada votes for the Liberal party, holding all other variables fixed;
- 0.833 is the coefficient for the dummy variable x_{BC}, where $x_{BC} = 0$ indicates the voter is not living in British Columbia, and $x_{BC} = 1$ if the voter lives in British Columbia. Therefore 0.833 represents that the log odds of the probability that a voter lives in British Columbia votes for the Liberal Party is expected to change by 0.833 compared with the log odds of the probability that a voter not living in British Columbia votes for the Liberal Party, holding all other variables fixed.
- The other coefficients can be interpreted in similar ways as above.

Table 2: Coefficients Table

Predictor	Liberal	Conservative	NDP	Bloc Quebecois	Green	People
intercept	-2.212	1.234	-0.344	-21.427	-3.218	-3.231
$x_{age18-24}$	Reference	Reference	Reference	Reference	Reference	Reference
$x_{age25-34}$			-0.538			
$x_{age35-44}$			-0.904			
$x_{age45-54}$			-1.427			
x_{age55+}			-1.649			
x_{Female}	Reference	Refernce	Reference	Reference	Reference	Reference
x_{Male}		0.662	-0.49		-0.236	
x_{Canada}	Reference	Reference	Reference	Reference	Reference	
x_{Other}	0.613			-1.759		-16.239
x_{AB}	Reference	Reference	Reference	Reference	Reference	Reference
x_{BC}	0.833	-1.971		0.111	1.566	
x_{MB}	1.011	-1.124		-0.047	0.725	
x_{NB}	1.235	-1.438		-0.067	1.653	
x_{NL}	1.669	-2.308		-0.049	-0.453	
x_{NS}	1.53	-1.94		-0.089	1.261	
x_{ON}	1.523	-1.645		0.073	0.541	
x_{PE}	1.578	-1.799		-0.103	1.524	
x_{QC}	1.322	-2.504		20.478	0.515	
x_{SK}	0.232	-0.416		-0.09	-0.341	
$x_{Married}$	Refernce	Reference	Reference	Reference	Reference	Reference
$x_{Not-married}$		-0.621			0.302	-0.777
$x_{No-university}$	Reference	Reference	Reference	Reference	Reference	Reference
$x_{University}$	0.496	-0.71			0.331	-1.49

The last step towards predicting the proportion of voters for the parties was completed in the post stratification processes. By dividing cells and estimating the log odds of the probability of voting for each one of the parties in each cell, the cell-level probabilities and proportions were calculated and aggregated using the cell size as weights. Finally, these proportions were summed up and divided by the total census population size to give us the population-level estimates of the voting proportions. According to Table 3, the Conservatives Party was predicted to have the highest proportion of voters, 36.8%, than other parties based on the post-stratification analysis using a logistic regression model with sex, province, marital status and education as predictors. It is worth mentioning that the sum of the following six proportions was approximately 1, which means the assumption that everyone had voted was satisfied and a almost perfect turnout rate was achieved.

Table 3: Predicted Proportion of Votes Table

Party	Liberal	Conservative	NDP	Bloc Quebecois	Green	People's
Proportion of Votes	33.1%	36.8%	14.8%	4.6%	9.7%	1.7%

Discussion

Summary

In this study, a survey data and a census data were accessed to build up six different models and they were used to retrospectively predict the 2019 Canadian Federal Election result using post stratification techniques, with an important assumption that the turnout rate being one. The first step was to clean up the data and standardized variable categories in both datasets. Some of the summary plots are shown in Figure 3; to have a clear view of the summary statistic of the data, refer to Table 1 and Figure 1 for more information. Then the "unsure" voters in the survey dataset were removed to avoid biased results and to fulfill the assumption that everyone had voted. Next, six frequentist logistic regression models were established for each party using the survey data from cesR package, and they were validated using ROC Curves and respective AUC values. Finally, the census data (GSS) was post-stratified into smaller cells and the population-level proportion was aggregated from cell-level estimates to predict the overall popular vote of the 2019 Canadian Federal Election. It was estimated that 36.8% of the voters would vote for the Conservative Party, which was the highest one among the six parties.

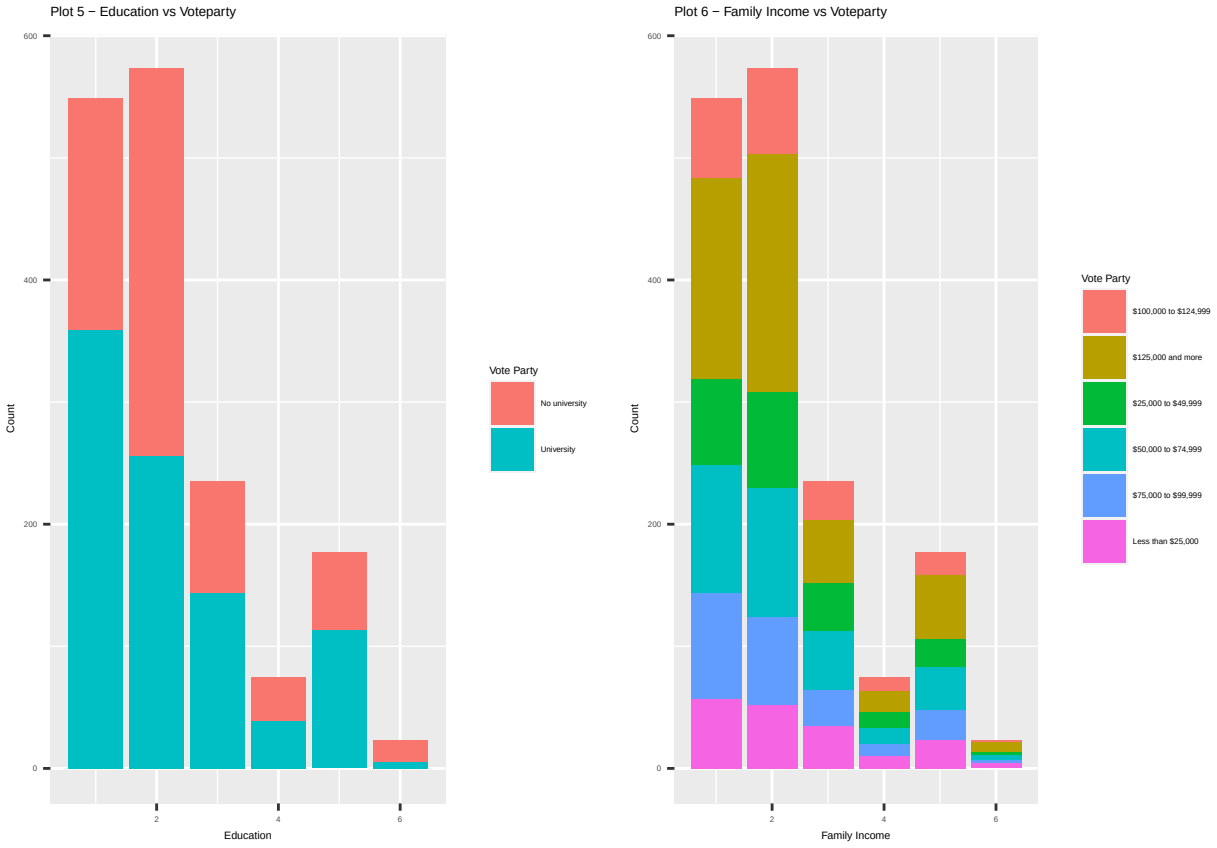

Figure 3: Barplots of two important variables that were considered in the models

Conclusion

Based on the result of the analysis, we can conclude that if everyone had voted in the 2019 Canadian Federal Election, then the Conservative Party would possibly be the winning party with 36.8% of total votes instead of the currently governing Liberal Party, which only had 33.1% of the votes. In addition, it was evident that predictors such as province, education, marital status and age groups would significantly affect the voters' choices, for example, only Quebec citizens would vote for Party Bloc Quebecois. (Figure 4) Therefore, if the turnout was higher and everyone had voted for a party, then the final result could be completely different and the Conservative Party would likely win the 2019 election.

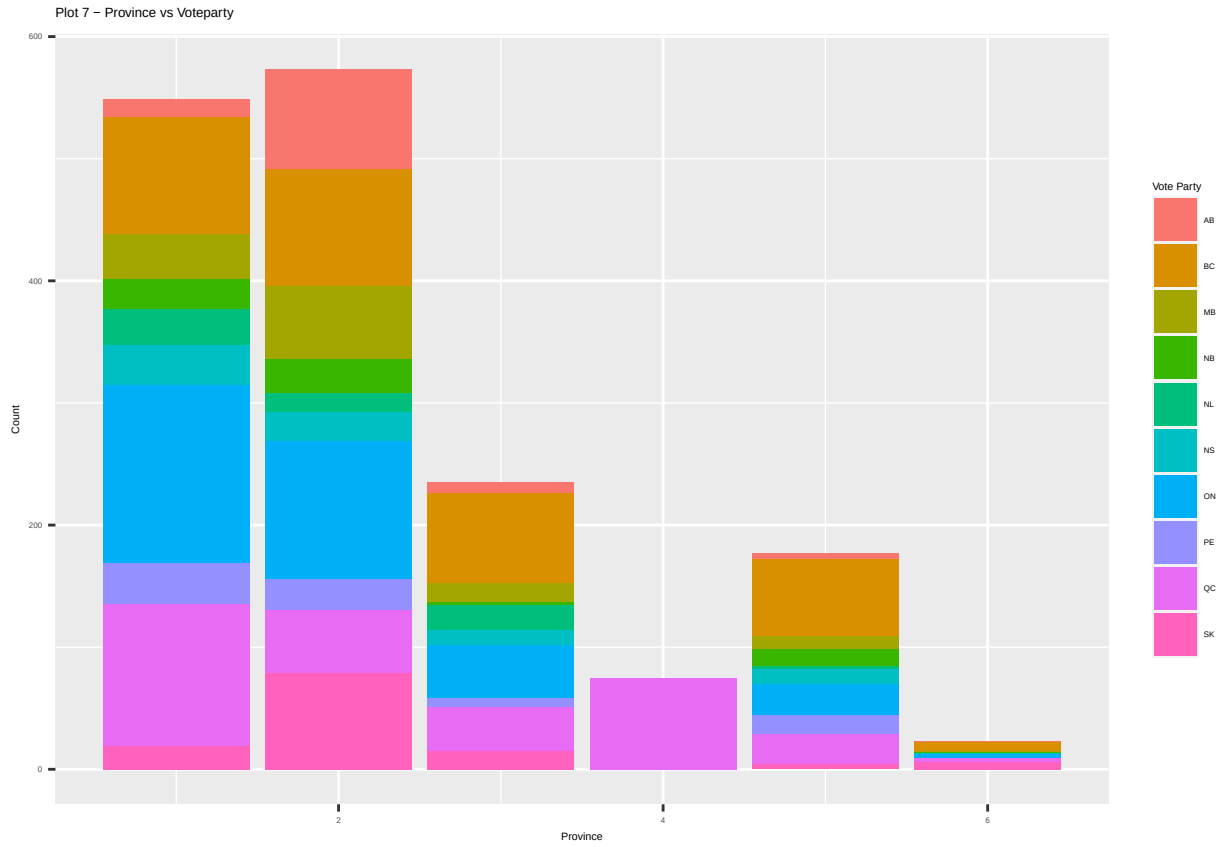

Figure 4: Barplots of variable province showing different voting choices between different groups

We can also get other insights about the relationship between the winner party and the turnout rate. This analysis indicates that, if this survey data can represent all Canadians and assume all of them had voted in the 2019 election, the Conservative Party could be the one that better represents people's wills and expectations in Canada. Even though the Liberal Party was the winner, getting the highest number of votes did not necessarily imply that most Canadians supports the Liberal Party; it could just be the case that, among all Canadians who voted, more of them supported the Liberals. There could be more supporters of the Conservative Party or other parties among those people who did not vote, and this is where the turnout rate turns out to be crucial to the final outcome of the election.

Weakness

Some drawbacks of this analysis are worth mentioning, for example, the census data used for post-stratification was from 2017, while the survey data was obtained in 2019. Many discordance could occur during this three-year period, such as how people's marital status changes and how their voting preference changes. Thus the accuracy of the predictions might be influenced by the inconsistencies. Secondly, many variables in the two datasets are quite different, limiting the number of variables that could be accounted for in the models. This makes certain important variables that could potentially change the outcome being neglected because they are missing in one of the datasets. For instance, vote choices in 2015 were recorded in the survey data but not in the census data while important inferences could be made on this variable. Thirdly, people who did not have clear intentions for the vote were filtered out in the survey dataset, but they might have voted in reality and this could significantly influence the result of the election. Besides that, removing indecisive interviewees made the size of the survey data even smaller, for example, the sharp corner in the ROC Curve for the Bloc Quebecois Party could be caused by the scarcity of data points. Any of these aspects could lead to non-representative data and biased results.

Next Step

For the next step, it can be a good way to improve these models by collecting voters' demographic information and the candidate party they actually voted for in the 2019 Canadian Federal Election, and comparing the results with the actual election results. In that way, I could rebuild the logistic regression models for a post hoc analysis; it is also an appropriate approach to use multilevel regression models to emphasize the differences between groups. For example, how Canadians from different provinces can differ in terms of their voting choices. Furthermore, more researches could be made to investigate what other features tend to influence Canadian's political preferences, which allows me to refine my models in order to generate more precise predictions.

References

1. cesR. Paul A. Hodgetts and Rohan Alexander (2020). cesR: Access the CES Datasets a Little Easier.. R package version 0.1.0.

2. cesR: An R package for the Canadian Election Study, SocArvix Papers. https://osf.io/preprints/socarxiv/a29h

3. Access the Canadian Election Study Datasets a Little Easier. https://hodgettsp.github.io/cesR/

4. 2017 General Social Survey (GSS): Families Cycle 31. https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/index.htm

5. Poststratification, Stan User;s Guide, https://mc-stan.org/docs/2_23/stan-users-guide/poststratification.html

6. Tidyverse. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

7. Visdat. Tierney N (2017). "visdat: Visualising Whole Data Frames." *JOSS*, *2*(16), 355. doi: 10.21105/joss.00355 (URL: https://doi.org/10.21105/joss.00355), <URL: http://dx.doi.org/10.21105/joss.00355>.

8. Skimr. Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2020). skimr: Compact and Flexible Summaries of Data. R package version 2.1.2. https://CRAN.R-project.org/package=skimr

9. Lme4. Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

10. pROC. Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77 http://www.biomedcentral.com/1471-2105/12/77/

11. ggplot2. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

12. ggpubr. Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

13. Reasons for not voting in the federal election„ Statistics Canada, https://www150.statcan.gc.ca/n1/daily-quotidien/200226/dq200226b-eng.htm

14. "gss_cleaning.R" Rohan Alexander and Sam Caetano, 2020, https://q.utoronto.ca/courses/184060/files/9422740?module_item_id=1867317