A Classification Approach to Examine "Would the Liberal Party Still be in Power Should Everyone Voted in 2019"

Weizhi Guo https://github.com/jayjes/304Final-project.git

22nd December, 2020

Abstract

Would Justin Trudeau and his Liberal Party win the 2019 election if every single eligible Canadian voter has voted in 2019? We attempt to explore the this question by formulating it as a classification problem based on a post 2019 election survey study and 2016's census data.

Keywords: 2019 Canadian Election, Classification, Logistic Regression

Introduction

In the past two year, Justin Trudeau and his Liberal Party was heavily criticized for corruption[3], their approach of handling international relations[2], their response to controlling the spread of the novel coronavirous[13], and numerous of other domestic and international affairs. It's only normal that one is curious about what would have happened if Andrew Scheer and the Conservative Party, or even Jagmeet Singh and the NDP has won the election. Would Canada have been in a better position if the country is under a different leadership. It would be difficult to obtain evidence to make a conclusive statement on this matter. One can, nonetheless, run simulations and make predictions on, for example, Canada's economic status if the country is under a different governing party. It would be a very difficult task both mathematically and computationally as there are thousands of variables to investigate and control for. One always, however, has to start with the question "what would have happened it the election outcomes were different", and a way to beginning this investigating is to find out "what would have happened if everyone voted in 2019". In this article, we will be employing multilevel regression and post-stratification method along with a logistic regression to classify voters' voting choices. The variables we will be controlling for are voters' age, gender, education level, and region where they casted their votes. Once the model is established, we will be using it to predict the 2019 election result as if everyone has voted.

Methodology

We will be utilizing the Canadian Election Study[11] (CES)'s 2019 post-election survey data as our training data. As per the documentation mentions, the dataset consists of 37822 observations. The gender distribution of the study population was targeted at 50% male and 50% female participants. The study population were targeted to include 28% of people who were of the age from 18-34, 33% from 35-54, and 39% of age 55% and above. 80% of French Canadian and 20% of English speaking Canadian were also targeted at the study population with Quebec, 10% French Canadian within the Atlantic provinces (Newfoundland and Labrador, New Brunswick, Nova Scotia, Prince Edward Island), and 10% French Canadian nationally. These were ambitious targets. Also, observational studies rarely attain their goals of sampling from the target population. We can expect some measurement bias in estimation from the CES data. Furthermore, as an observational study, the CES data was highly likely to be subjected selection bias. As a result, we will be utilizing Statistics Canada's *Highest level of educational attainment (general) by sex and selected age groups* census data[5] for post stratification and prediction. One would, normally, consider census data as a more reliable population

data source even though one cannot completely rule out the possibility of measurement or selection bias. The census data we will be using was from 2016, which was the latest census data we could obtain. We settled on this dataset because people's level of education was known to effect voting behaviours[1].

We selected `cps19_votechoice`, `cps19_gender`, `cps19_province`, `cps19_education`, `cps19_age` from the CES data. These variables correspond to the voter's vote choice, gender, location, their education level, and age. Upon close inspection of the dataset, we saw 6258 missing values in the `cps19_votechoice` column and 0 missing values in the rest of the variables. As, 6258 observations were not a big proportion of the 37822 study populations. We will be deleting these observations. We will also be collapsing some factors of the voter's response. Specifically, we will be collapsing the "Another party (please specify)", and "Don't know/ Prefer not to answer" into a `Other` category. We will also discretize people's age into bins `[18,34)`, `[34,55)`, and `[55,120)`. Finally, we will only consider the education groups`>BA`, `<BA`, and `college`, which correspond to people with a bachelor's degree or higher, people without a college or professional degree, and people with a college or professional degree, respectively. The educational census data from Statistics Canada will also be cleansed in a similar manner. The census data will be used to calculate the post-stratification proportions for each predictor, and these proportions will be used to predict the 2019 election outcome.

In Figure 1, we examine if the CES dataset's sampling goal was achieved with by visualizing their frequency

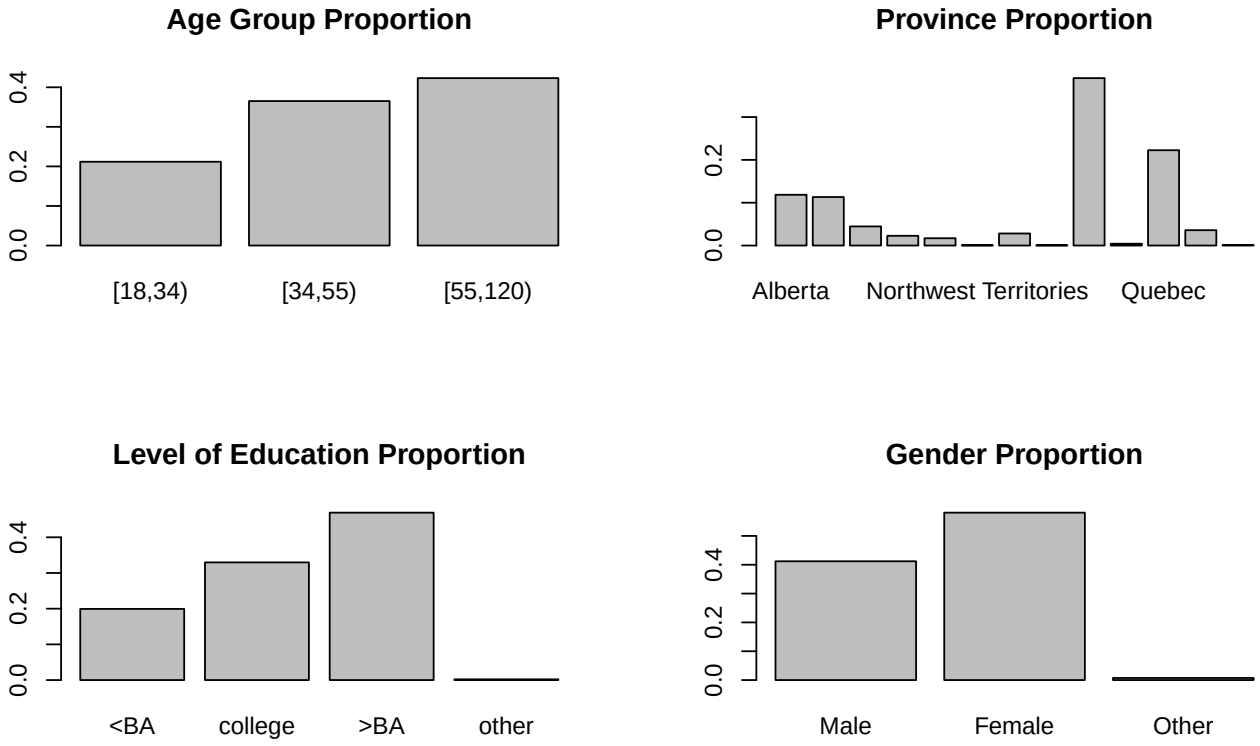

Figure 1: Proportion of the Study Sample

We see that the age group's objective was roughly achieved, but the population from each province or territory was clearly not evenly sampled. The sample from Ontario dominates most of the study population. Quebec comes next. One should keep in mind that some provinces or territories such as the New Brunswick and Northwest Territories would be under-represented. For level of education and gender, no one obvious category was over or under sampled.

For the mathematical formulation of the model, we will be using multinomial logistic regression in the classical framework to classify voters' voting choice based on location, age groups, education level, and gender. For modeling The mathematical formulation is as follow

$$\text{logit } \mathbb{P}(Y_i = j) = \beta_0 + \beta_1 \times \text{location} + \beta_2 \times \text{age} + \beta_3 \times \text{education} + \beta_4 \times \text{gender} \tag{1}$$

for $i \in [1, 31564]$, and j is the enumeration of people's voting choice, and $j = 1, 2, 3, 4, 5, 6, 7$, representing the Liberal, Conservative, NDP, Bloc Quebecois, Green Party, People's Party, and Other, respectively. We use logistic regression in the classical framework because all our predictors are of discrete nature. We lack additional knowledge in terms of their prior distributions. As a preliminary estimation, however, a logistic regression model in the classical framework will suffice.

Results

We first evaluate the model's performance with the training data. We obtain the following confusion matrix.

	Liberal	Conservative	NDP	Bloc Québécois	Green Party	People's Party	Other
Liberal	5873	3653	2216	863	1462	294	2858
Conservative	2064	4284	1380	0	681	225	1547
NDP	422	385	590	2	204	46	201
Bloc Québécois	346	246	51	388	37	16	259
Green Party	13	9	12	0	15	4	14
People's Party	0	0	0	0	0	0	0
Other	231	136	79	151	57	20	230

This model yielded 0.36054 training error. We will further discuss the implication of this model under the discussion section.

We will proceed to predict with the post-stratified data next.

In Figure 2, we aggregated the people's voting intention by province and obtained that the Conservative party would have won in Alberta, BC, Manitoba, and Saskatchewan. The Liberal Party, on the other hand, would have won the rest of the provinces and territories. Bloc Québécois only got support in Quebec. The Green Party had relative large support from Prince Edward Island.

In Figure 3, we aggregated the data by age groups. We saw younger people tend to vote for parties other than the Conservative and the Liberal Party. But the Liberal Party still dominate across all age groups, despite that the older the population, the more likely they vote for the Conservative Party.

We considered the education levels in Figure 4. People without a Bachelor's degree is more likely to vote for the Conservative Party, while the people with a Bachelor or higher degree tend to vote for the Liberal Party.

Finally, in Figure 5, aggregating by gender, men are more much more likely to cast their votes for the conservative Party. Both genders have equal tendency to vote for the Liberal Party.

Overall, based on the model we have after post-stratification and assume that everyone has voted, Justin Trudeau and his Liberal Party would still won the 2019 election, and the Conservative Party would still be in the runner-up.

Discussion

For this study we used the CES dataset. As we said in the methodology section, the CES was an observational study, and the response were collected via a survey. Several problems arise with this format of study. First of all, the volunteer bias and recall bias can be potential problems. Those who consented to participate in the study might be those people who are politically active. Their counterparts are under-represented in the CES study. Furthermore, when filling out surveys, people might have difficulty filling in the most accurate information. Measurement bias might also complicate the reliability of the study. As per the CES documentation mentions, the practitioners had specific sampling goals, but their goals might be difficult to achieve, and if the goal was not attained, certain population's voting information might be less accurate.

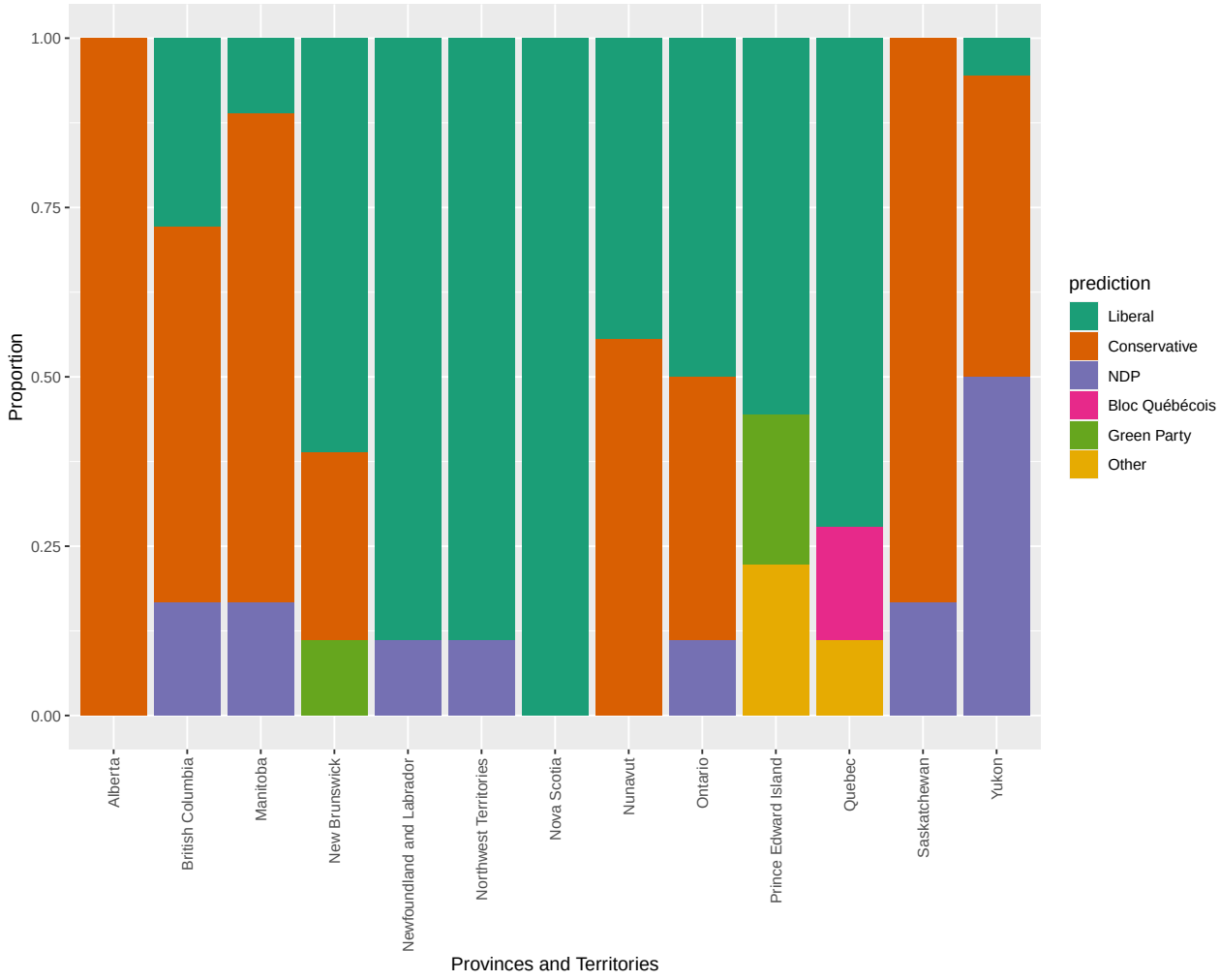

Figure 2: Voting Choice Aggregated by Province

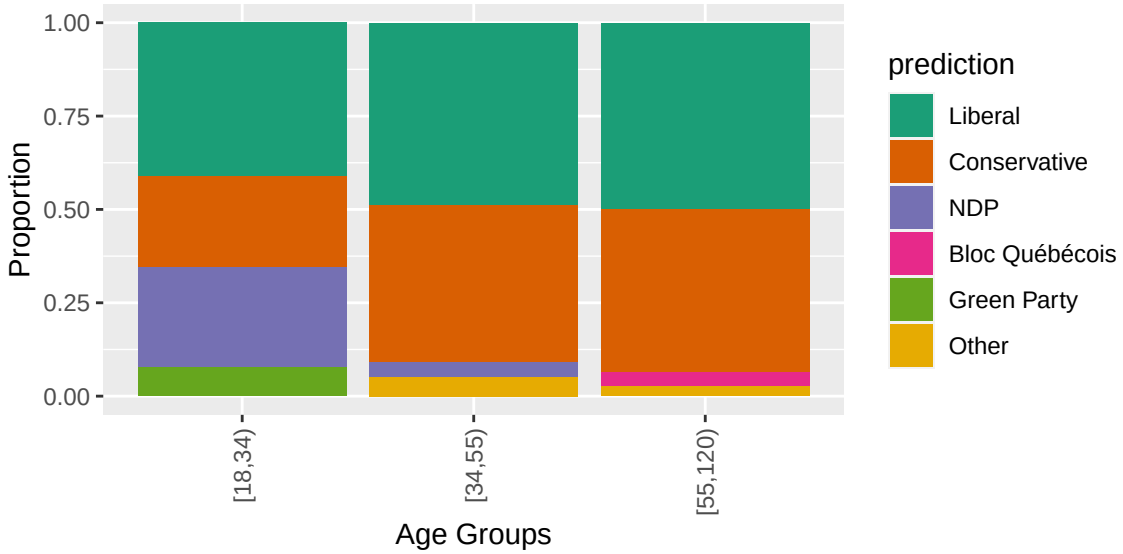

Figure 3: Voting Choice Aggregated by Age

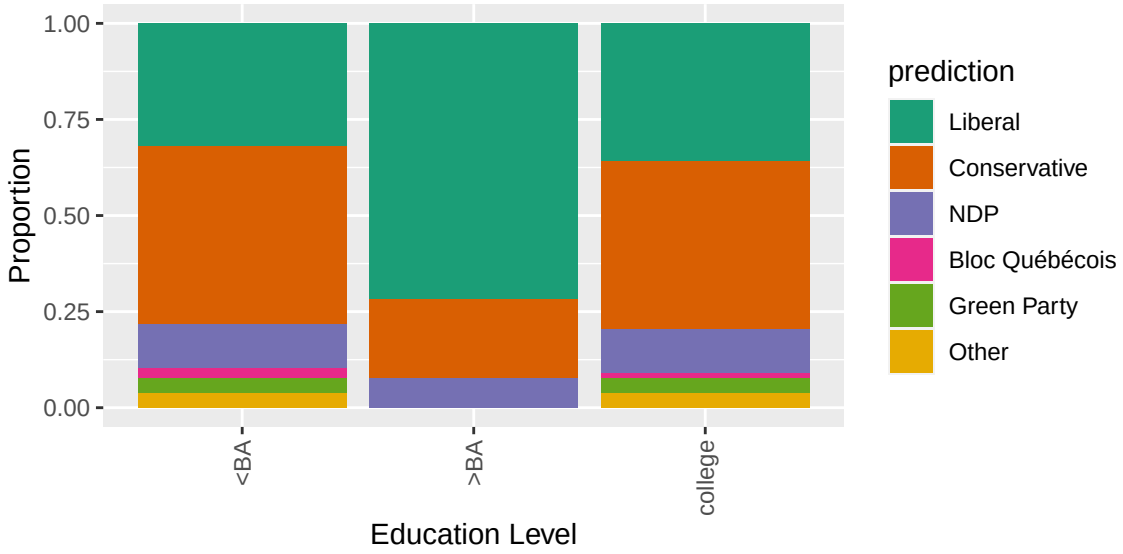

Figure 4: Voting Choice Aggregated by Education Level

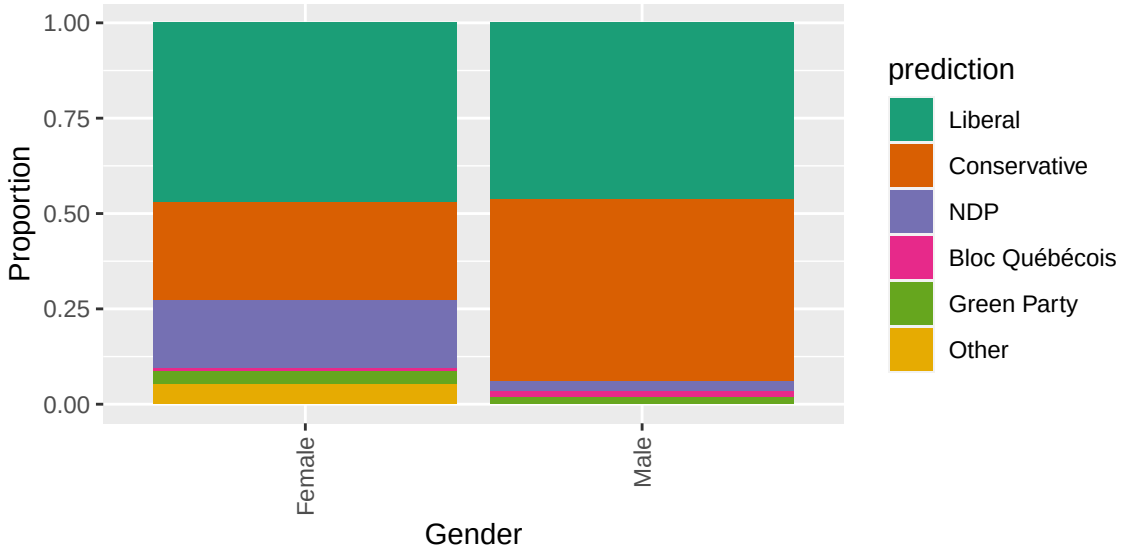

Figure 5: Voting Choice Aggregated by Gender

Additionally, we noticed that the documentation did not mention anything about the Native Canadian population. We were unsure if the practitioners neglected the Native Canadians or they implicitly assumed they were able to obtain an adequate sample that represents the Native Canadians. We briefly examined if the outlined goal was achieved and concluded that the study population were dominated with samples from Ontario and Quebec. We expect the mentioned problems with the CES dataset to cause bias in the model as we did not use the census data to train our model. We hope that post-stratification would rectify the bias as we said census data are generally more reliable. Future investigations can be done to investigate if post-stratification actually lower the bias from a biased training dataset and model.

We did not perform variable selection in this article. We chose the "most relevant" ones based on common sense. Some valuable information may offer more insights into people's voting intention. Future works can perform more rigorous and systematic method of predictors selection. We could take into account for the fact that the language spoken could potentially affect people's voting intention. Interactions between predictors could be further investigated as well.

The prediction results divulged that the Liberal Party would still win the 2019 election assuming everyone has voted. We will not be discussing and analyzing any political implication in this article. Instead we will discuss about the model. Readers might have noticed that the logistic regression model was performing poorly in terms of the predicting the voter's intention. The accuracy was quite low. The low accuracy might be a result of few causes. Firstly, the CES dataset might not suit the format of logistic regression. For example, the voters might not form clear clusters with unambiguous hyper-plane boundaries. Even though we have collapsed some factors, the number of factors in `province` is still relatively high. Further investigations can be done to select a suitable model for the CES dataset. Non-parametric models such as the tree-based models or boosting might work better than logistic regression. Another reason might be that we did not fully utilized the available information for the logistic regression model. Election results are well studied subjects. One can construct a suitable prior distribution from a previous study on our predictors and utilize one of the Bayesian classification methods to, hopefully, improve prediction accuracy.

Conclusion

In this article, we utilized multinomial logistic regression and post-stratification to examine the question "what if everyone has voted in the 2019 election". We gathered training data from CES and educational level census data and chose age groups, educational level, gender, and the voter's residing province territory as predictors. We cleansed the two datasets into compatible formats. For the CES dataset, in particular, we aimed to follow CES's documentation and organized voters' age into bins. We also organized the educational levels into people with a Bachelor's degree or a profession degree or without the two said degrees. Collapsing the age and educations not only helped to reduce the number factors, the method combines factors with very few observations. This would minimize the number of outliers, which is generally consider an obstacle for regression models. Despite the fact that our prediction accuracy was low, one can get a preliminary idea on what would have happened if every Canadian citizen voted in 2019, controlling for age, gender, level of education, and voters' residing province. The election outcome is still the same as it was today.

Appendix A

The following are the post-stratified predicted datasets aggregated by each predictor

Table 1: Voting Intention Group by Province/Territory

province	prediction	n	freq
Alberta	Conservative	18	1.0000000
British Columbia	Liberal	5	0.2777778
British Columbia	Conservative	10	0.5555556
British Columbia	NDP	3	0.1666667
Manitoba	Liberal	2	0.1111111
Manitoba	Conservative	13	0.7222222
Manitoba	NDP	3	0.1666667
New Brunswick	Liberal	11	0.6111111
New Brunswick	Conservative	5	0.2777778
New Brunswick	Green Party	2	0.1111111
Newfoundland and Labrador	Liberal	16	0.8888889
Newfoundland and Labrador	NDP	2	0.1111111
Northwest Territories	Liberal	16	0.8888889
Northwest Territories	NDP	2	0.1111111
Nova Scotia	Liberal	18	1.0000000
Nunavut	Liberal	8	0.4444444
Nunavut	Conservative	10	0.5555556
Ontario	Liberal	9	0.5000000
Ontario	Conservative	7	0.3888889
Ontario	NDP	2	0.1111111
Prince Edward Island	Liberal	10	0.5555556
Prince Edward Island	Green Party	4	0.2222222
Prince Edward Island	Other	4	0.2222222
Quebec	Liberal	13	0.7222222
Quebec	Bloc Québécois	3	0.1666667
Quebec	Other	2	0.1111111
Saskatchewan	Conservative	15	0.8333333
Saskatchewan	NDP	3	0.1666667
Yukon	Liberal	1	0.0555556
Yukon	Conservative	8	0.4444444
Yukon	NDP	9	0.5000000

Table 2: Voting Intention Group by Age

age_group	prediction	n	freq
[18,34)	Liberal	32	0.4102564
[18,34)	Conservative	19	0.2435897
[18,34)	NDP	21	0.2692308
[18,34)	Green Party	6	0.0769231
[34,55)	Liberal	38	0.4871795
[34,55)	Conservative	33	0.4230769
[34,55)	NDP	3	0.0384615
[34,55)	Other	4	0.0512821
[55,120)	Liberal	39	0.5000000
[55,120)	Conservative	34	0.4358974

age_group	prediction	n	freq
[55,120)	Bloc Québécois	3	0.0384615
[55,120)	Other	2	0.0256410

Table 3: Voting Intention Group by Education

edu_group	prediction	n	freq
<BA	Liberal	25	0.3205128
<BA	Conservative	36	0.4615385
<BA	NDP	9	0.1153846
<BA	Bloc Québécois	2	0.0256410
<BA	Green Party	3	0.0384615
<BA	Other	3	0.0384615
>BA	Liberal	56	0.7179487
>BA	Conservative	16	0.2051282
>BA	NDP	6	0.0769231
college	Liberal	28	0.3589744
college	Conservative	34	0.4358974
college	NDP	9	0.1153846
college	Bloc Québécois	1	0.0128205
college	Green Party	3	0.0384615
college	Other	3	0.0384615

Table 4: Voting Intention Group by Gender

gender	prediction	n	freq
Female	Liberal	55	0.4700855
Female	Conservative	30	0.2564103
Female	NDP	21	0.1794872
Female	Bloc Québécois	1	0.0085470
Female	Green Party	4	0.0341880
Female	Other	6	0.0512821
Male	Liberal	54	0.4615385
Male	Conservative	56	0.4786325
Male	NDP	3	0.0256410
Male	Bloc Québécois	2	0.0170940
Male	Green Party	2	0.0170940

Appendix B

We used `nnet::multinom` to build our multinomial classification model. `nnet`'s implementation is based on feed-forward neural network and it was consider a more rigorous implementation than `glm` for multinomially models.

Reference

1. Archer, K. (1987). A simultaneous equation model of Canadian voting behaviour. Canadian Journal of Political Science/Revue canadienne de science politique, 20(3), 553-572.
2. Black, C. (2020, December 04). Conrad Black: The Liberal government's policy of self-impoverishment will hurt us all. Retrieved December 22, 2020, from https://nationalpost.com/opinion/conrad-black-the-liberal-governments-policy-of-self-impoverishment-will-hurt-us-all.
3. Blatchford, A. (2020, July 24). Trudeau strains to contain political scandal engulfing his family. Retrieved December 22, 2020, from https://www.politico.com/news/2020/07/24/trudeau-political-scandal-family-381002.
4. David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.3. https://CRAN.R-project.org/package=broom.
5. Statistics Canada, (2017, November 27). Education Highlight Tables, 2016 Census. Retrieved December 22, 2020, from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/edu-sco/index-eng.cfm.
6. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
7. JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2020). rmarkdown: Dynamic Documents for R. R package version 2.3. URL https://rmarkdown.rstudio.com.
8. Joseph Larmarange (2020). labelled: Manipulating Labelled Data. R package version 2.7.0. https://CRAN.R-project.org/package=labelled
9. Paul A. Hodgetts and Rohan Alexander (2020). cesR: Access the CES Datasets a Little Easier.. R package version 0.1.0.
10. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
11. Stephenson, Laura B., et al. 2019 Canadian Election Study - Online Survey. 1 May 2020. dataverse.harvard.edu, doi:10.7910/DVN/DUS88V.
12. Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.
13. Wherry, A. (2020, November 27). The Conservatives fire up a phoney war over the 'Great Reset' theory | CBC News. Retrieved December 22, 2020, from https://www.cbc.ca/news/politics/great-reset-trudeau-poilievre-otoole-pandemic-covid-1.5817973
14. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686.
15. Yihui Xie and J.J. Allaire and Garrett Grolemund (2018). R Markdown: The Definitive Guide. Chapman and Hall/CRC. ISBN 9781138359338. URL https://bookdown.org/yihui/rmarkdown.