The Election Result Would Not Change If All Ages Citizens Had Voted for the 2019 Canadian Federal Election

Qing Li(1005148010)

2020/12/22

Code and data supporting this analysis is available at: https://github.com/Alicia-LQ/QL

Abstract

The objective of this report is to predict which party will win the election if all ages citizens in Canada had voted for the 2019 Canadian Federal Election. Models will be built based on the online survey results of Canadian Election Study in 2019. Next, apply the model to the 2016 census data to find which party will have the highest votes. Finally, comparing the predicted result with the actual election result of 2019 to find if there will exist any difference.

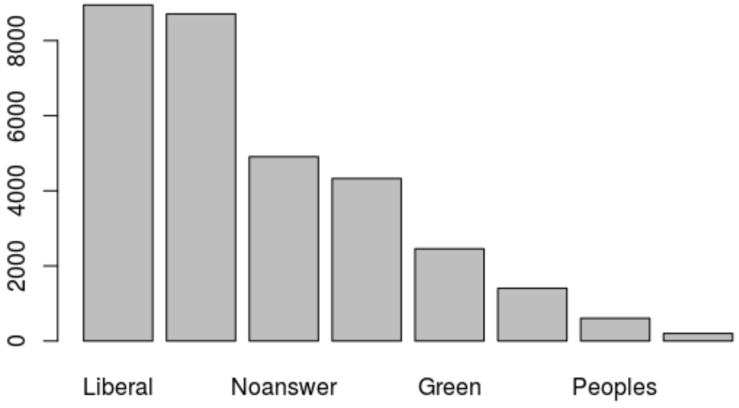
Keywords: 2019 Canadian Federal Election, multi-class classification, multilevel regression with post-stratification(MRP), logistic regression, ROC, AUC, softmax

Introduction

The survey result of 2019 Canadian Election Study has more than 37000 survey answers on over 600 questions. In our project, the target population is all Canadian citizens; the sample population consists of 37822 respondents.

In the model, "age","province", "education" and "gender" will be considered as independent variables and "vote choice" will be the only response variable. Additionally, eight vote choices are included in the outcome variable and their distributions are shown in Figure 1. It reveals that the majority chose "Liberal Party"; "Conservative Party" was the second popular choice; "People's Party" had the lowest vote. Next, using the mentioned variables to build a logistic regression, and then implement this model to the 2016 census data to predict if all ages citizens had voted then what would happen to the election result.

Figure 1

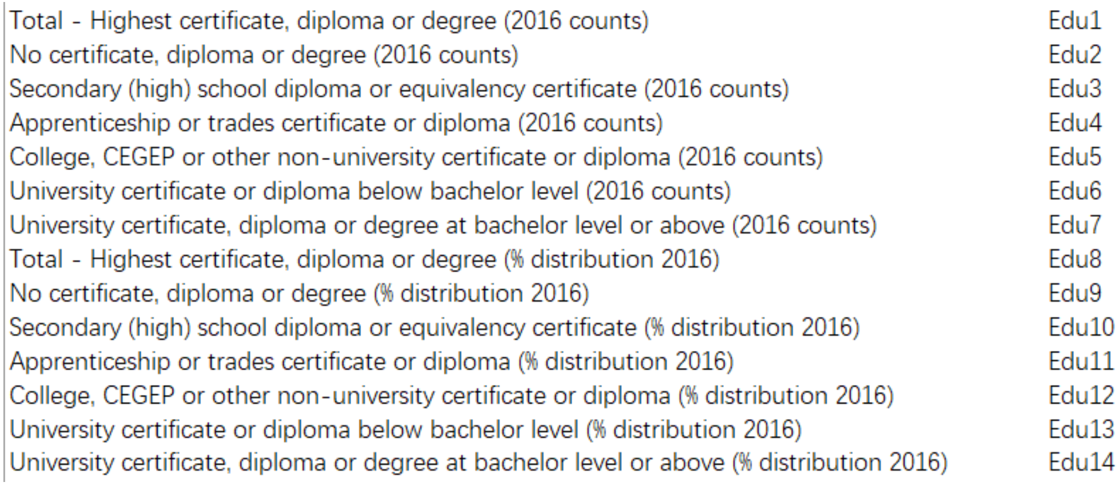


Data

The survey data is obtained from Canadian Election Study, and the census data is obtained from Statistics Canada. To be consistent with the census data, variable "age" in survey data will be divided into five groups, "education" will be divided into six groups, "gender" will contain three groups, and "province" will keep the same. For the outcome variable, "Liberal Party", "Green Party", "Conservative Party", "ndp" and "Bloc Qu" will keep the same; the rest of the eight choices will be combined together, and they will be called as "Others". In addition, the data contain lots of non-response, and the reason why people refused to vote for a party was that they did not care about the election result, or they did not trust any of the parties.

Moreover, the census data contain variables of provinces, age, sex, and education level. Education levels from the census data are renamed by "Edu*", and Table 1 refers to the variables mapping.

Table 1: Variables Mapping



Total - Highest certificate, diploma or degree (2016 counts)	Edu1
No certificate, diploma or degree (2016 counts)	Edu2
Secondary (high) school diploma or equivalency certificate (2016 counts)	Edu3
Apprenticeship or trades certificate or diploma (2016 counts)	Edu4
College, CEGEP or other non-university certificate or diploma (2016 counts)	Edu5
University certificate or diploma below bachelor level (2016 counts)	Edu6
University certificate, diploma or degree at bachelor level or above (2016 counts)	Edu7
Total - Highest certificate, diploma or degree (% distribution 2016)	Edu8
No certificate, diploma or degree (% distribution 2016)	Edu9
Secondary (high) school diploma or equivalency certificate (% distribution 2016)	Edu10
Apprenticeship or trades certificate or diploma (% distribution 2016)	Edu11
College, CEGEP or other non-university certificate or diploma (% distribution 2016)	Edu12
University certificate or diploma below bachelor level (% distribution 2016)	Edu13
University certificate, diploma or degree at bachelor level or above (% distribution 2016)	Edu14

Model

The survey data consists of a lot of multi-choice questions, and we would like to transfer them to multi-binary model by using one v.s rest method[1]. To explain, if we have n levels in outcome, then n outcome will be obtained. For each vote choice, it would be a binary question. In this project, the first vote choice would be choosing "Liberal Party" or not, and other vote choice works in the same way.

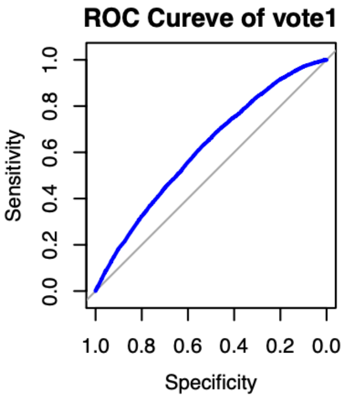
For binary questions, logistic regression model[5] will be trained by using the equation below. In addition, categorical variables with n levels will contain $n-1$ dummy variables. Next, each choice will train one logistic regression model, and then implement the model to census data to obtain the predicted probability of choosing the corresponding party.

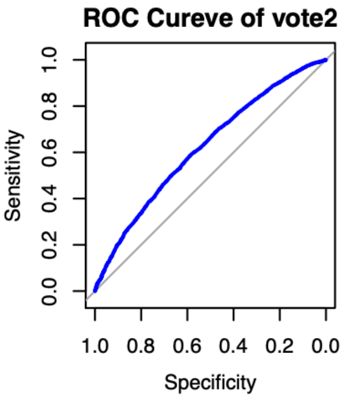
$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 age + \beta_2 gender + \beta_3 education + \beta_4 province$$

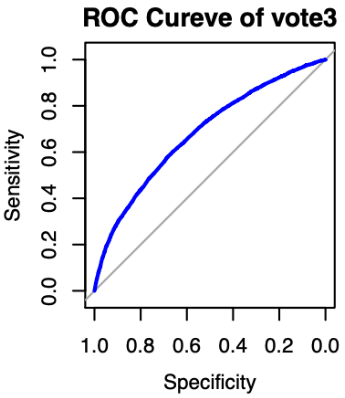
After that, one combined features group will have six predicted probability on different party. Then, we use softmax method[8] to convert the probability into a range of [0~1]. The formula of softmax method is shown below.

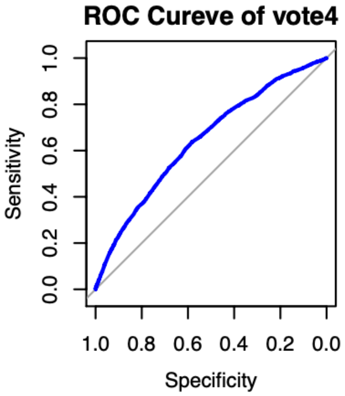
$$p_i' = \frac{p_i}{\sum(p_i)}$$

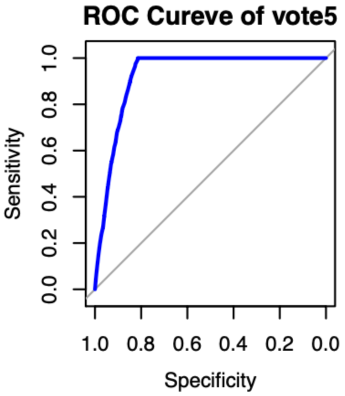
Next, using the ROC curve[3] to see how well the model is. Specifically, the ROC curve is "a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive." The area under ROC curve is called AUC. The larger the AUC is, the better the model will be. For our six binary models, the six corresponding ROC curves are shown below. We can see that the areas are large there, in other words, the AUC is large. Therefore, we can conclude that our model is built properly.

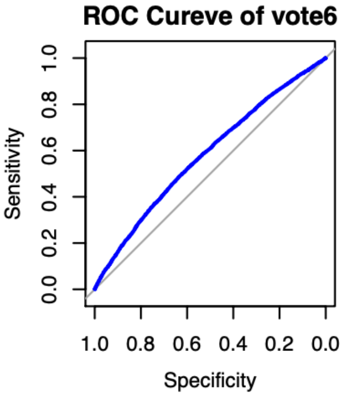






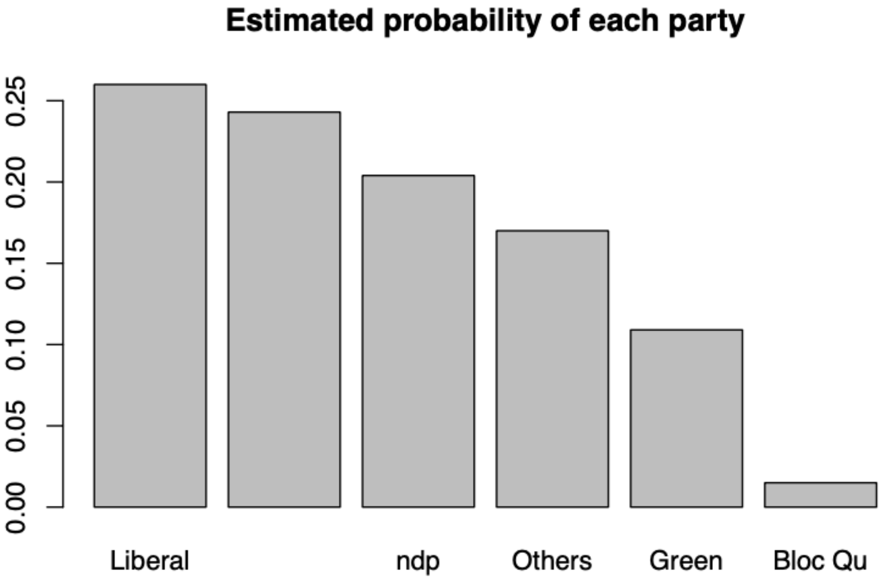






Results

One of the model results is shown below. It reflects that all features are significant when significance level is set as 0.1, but if we set the significance level to 0.05, then variable "education" would not be significant any more. Moreover, for the coefficient of each variable, let's take "gender" as the example, there are three levels in variable “gender”(Male, Female, Others). Specifically, female level is treated as the reference; the other two will compare with it. In addition, the coefficient of "genderMale" is -0.34731, it means that the odds ratio is $exp(-0.34731) = 0.7066$.



As we can see from the above plot, "Liberal Party" has the biggest predicted probability, which is approximately 0.26. "Conservative Party" is the second highest. Thus, we suggest that Liberal Party would win the election if all ages citizens in Canada had voted for the 2019 Canadian Federal Election. Our predicted result is exactly same as the actual election result of 2019 Canadian Federal Election[6].

Discussion

In this analysis, the online survey results of Canadian Election Study in 2019 is used as sample to build our prediction models. The survey contains multi-choice questions, and we have transferred them to binary questions, and we use softmax method to generate the predicted probability. According to our predicted result, we can conclude that the election result would not change if all ages citizens were considered. In other words, the winner will always be Liberal Party. However, the gap between Liberal Party and Conservative Party is not as big as the actual election result has showed. There is a change that the Conservative Party would predicted to win if we change the predictor variables or census data.

Possible Future Improvements

Here are some future improvements for making such prediction. Firstly, we can design and implement a better vote survey with more questions and try to get as many respondents to take the survey. Secondly, applying some other models such as decision tree to predict a more accurate result. Finally, we can add more predictor variables since other factors would influence the vote intention.

References

[1] Brownlee, J. (2020, September 07). One-vs-Rest and One-vs-One for Multi-Class Classification. Retrieved December 22, 2020, from https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/

[2] Canada Federal Elections. (1970, December 01). Retrieved December 22, 2020, from https://www.elections.ca/home.aspx

[3] Grace-Martin, K., Oehr, & Chamberlain, K. (2018, December 13). What Is an ROC Curve? Retrieved December 22, 2020, from https://www.theanalysisfactor.com/what-is-an-roc-curve/

[4] Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

[5] Logistic Regression. (n.d.). Retrieved December 22, 2020, from http://www.cookbook-r.com/Statistical_analysis/Logistic_regression/.

[6] Newton, P. (2019, October 22). Buckle up -- Canada's election will be a cliffhanger. Retrieved December 22, 2020, from https://www.cnn.com/2019/10/20/world/canada-election-october-21-intl/index.html

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

[8] Softmax Function. (2019, May 17). Retrieved December 22, 2020, from https://deepai.org/machine-learning-glossary-and-terms/softmax-layer

[9] Statistical Modeling, Causal Inference, and Social Science. (n.d.). Retrieved December 22, 2020, from https://statmodeling.stat.columbia.edu/

[10] The element of statistic learning: Linear model for Classification. (n.d.). Retrieved December 22, 2020, from https://web.stanford.edu/~hastie/ElemStatLearn/

[11] 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/>