A stastical analysis of voter demographics for the Liberal Party of Canada

Artem Arutyunov

December 7, 2022

Introduction

A crucial set of information for any political strategist consists of voter demographics. A political party must know their current voter demographics to plan an effective campaign which will maximize voter behaviour in their favour.

In this presentation, we conduct a statistical analysis of voter demographics for the purpose of advising the *Liberal Party of Canada*. We investigate the following research questions:

1. **Are people who selected Liberal as their first choice to vote equally as likely to support more or less refugees?**
2. **Does gender influence the median rating of Justin Trudeau?**
3. **Can we predict voter behaviour based on traits such as age, gender, and marital status?**

Data Summary

We used the data collected from the Canadian Election Study in 2019 (`ces19`), which consists of $37,822$ responses from online and phone surveys.

Variables used:

1. Vote choice (first choice to vote for): will be used to filter out people who are not voting for the Liberal Party, and to determine voting preference.
2. Refugee (should Canada admit more refugees?): will be used to classify the opinion of the voters. If their answer is more refugees, then they are supporting the admission, if it is fewer refugees, then they are against the admission. Answers such as "same number of refugees" or "don't know/prefer not to", will be ignored because this opinion can not be classified as strictly positive or negative.

Research Question I

Political parties should pay attention to the opinion of voters on important discussions that are present in Canadian society. One such issue regards immigrants and refugees.

For this question, we examine whether Liberal voters are equally as likely to support or oppose policies regarding admitting more refugees into Canada.

Null hypothesis (H_0): The proportion of voters whose first choice is the Liberal party who support more refugees is 50%.

$$H_0 : p_{\text{more}} = 0.5$$

Alternative hypothesis (H_1): The proportion of voters whose first choice is the Liberal party who support more refugees is not 50%.

$$H_1 : p_{\text{more}} \neq 0.5$$

where p_{more} refers to the proportion of voters.

Statistical Methods

To test the null hypothesis, we conducted a series of simulations in R.

- To find the test statistic, the proportion of Liberal voters (i.e. first choice is Liberal) who support the admission of refugees was calculated ($\hat{p}_{\text{more}} = 0.4242$).
- To get a sampling distribution, we ran a test where we randomly sample a position ("support" or "oppose") with uniform probability for each voter. We repeated this test $10,000$ times. For each test, the proportion of positive answers was calculated.
- The p-value was found by calculating the proportion of values in the estimated sampling distribution that are more or as extreme as the test statistic.

Results & Conclusions

- From our randomization test, we computed a p-value of 0.
- Since $P = 0 < 0.01$, the probability of observing a value *at least as extreme* as the test statistic ($\hat{p}_{\text{more}} = 0.4242$) is *extremely* unlikely. We have *very strong* evidence against the null hypothesis.

Therefore, we can conclude that the proportion of liberal party voters who support the policy is **not** 50%. In other words, this means that it is not equally likely for Liberal voters to take a random stance on the refugee issue.

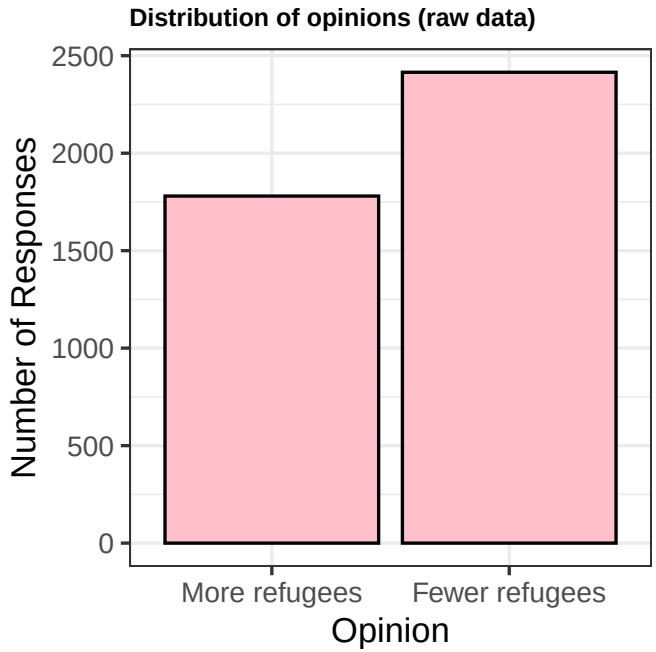

Figure 1: the distribution of the original data. As we can see, the distribution of the raw data is consistent with our conclusions from the randomization test (that $p_{\text{more}} \neq 0.5$).

Research Question II

For this question, we examine whether gender influences the median rating of party leader, Justin Trudeau?

Null hypothesis (H_0): There is no difference between the median rating of Justin Trudeau (`lead_rating_23`) between men and women:

$$H_0 : \text{median}_{\text{men}} - \text{median}_{\text{women}} = 0.$$

Alternative hypothesis (H_1): There is a difference between the median rating of Justin Trudeau (`lead_rating_23`) between men and women:

$$H_1 : \text{median}_{\text{men}} - \text{median}_{\text{women}} \neq 0.$$

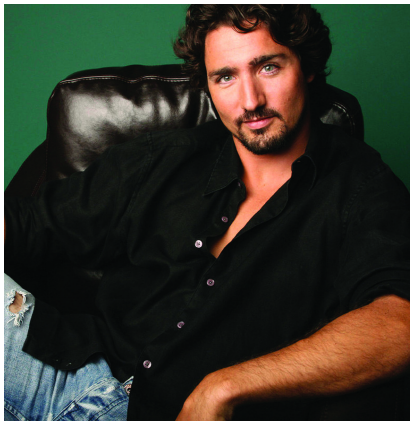

Figure 2: Justin Trudeau: Benevolent Party Leader and Prime Minister.

Figure 3: Opposing party leader, Andrew Scheer, drinks Neilson Dairy 2% skim milk.

Statistical Methods

To test the null hypothesis, we conducted a series of simulations in R (similar to the method used in part A).

- The test statistic is the difference between the median rating among men and the median rating among women:

$$\hat{\text{median}}_{\text{men}} - \hat{\text{median}}_{\text{women}} = 8.$$

- To get a sampling distribution, we ran a test where we randomly shuffle the gender labels for each observation with uniform probability. We repeated this test $10,000$ times. For each test, the difference between the median ratings was calculated.

Results & Conclusions

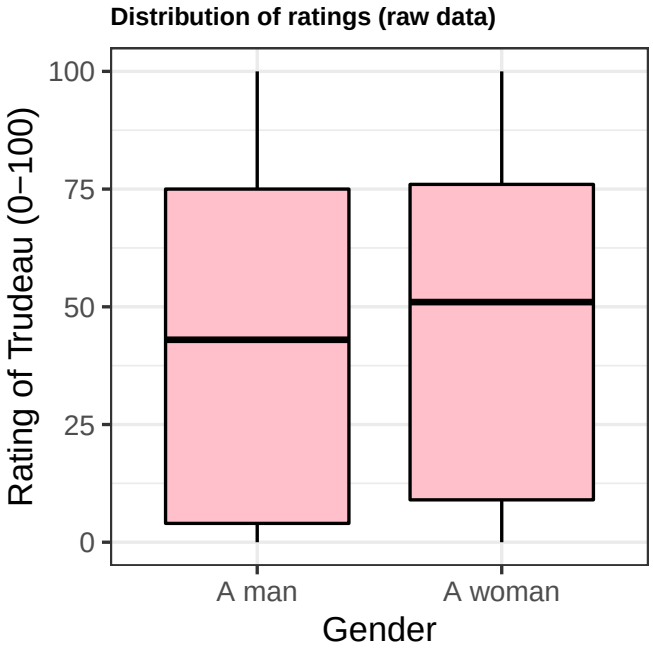

Figure 4: the distribution of ratings in the original data. The distribution of the raw data is consistent with our conclusions from the randomization test.

- From our randomization test, we computed a p-value of 0.
- Since $P = 0 < 0.01$, we have *very strong* evidence against the null hypothesis.
- Therefore, we can conclude there **is** a difference of median rating of Justin Trudeau between men and women.

These results show that when campaigning and advertising, we should consider the gender of potential voters to better target their expectations of party leaders. If our data is representative of the real-world, according to Figure 4, this may suggest that on the whole, women rate Trudeau higher than men.

Research Question III

Broadly speaking, are there certain traits that voters have that we can use to predict who they will vote for? Can age, gender, and marital status be used to predict a voter's choice? If we add more than those 3, will we have a clearer picture of predicted voter choice?

For this question, we will focus our attention on predicting whether a voter will vote for the top two parties of Canada: the Liberals or Conservatives.

Statistical Methods

We trained multiple classification models to predict voter choice (`votechoice`) based on the following variables:

- `age`, `gender`, `province`, `citizenship`, `bornin_canada`
- `sexuality`, `religion`, `marital`, `education`, `union`, `children`
- `groups_therm_1` to `groups_therm_5`
- `interest_gen_1`, `interest_elxn_1`
- opinions on spending: `spend_educ`, `spend_env`, etc...
- opinions on the economy: `econ_retro`, etc...
- opinions on immigrants/refugees: `imm`, `refugees`
- opinions on the government: `govt_confusing`, `govt_say`, and `lib_promises`.

Why so much data?

To accurately train a classification model, we need sufficient predictors to distinguish voters of different parties.

Statistical Methods (cont'd.)

The classification models were compared with one another on the basis of accuracy to find the most accurate model. We split the data into a training and test set (80/20), and then used them to train and evaluate the following models:

- **Classification tree**: by asking TRUE/FALSE questions about the predictors, we can narrow down the set of possible outputs, and if we keep doing this, we will converge to a single answer. Formally, each 'question' is called a binary split.
- **Neural Network**: a simplified model of the brain; mimics its structure with the hopes of finding relationships in the predictors (similar to that of a human).
- **Support Vector Machine**: an approach for "dividing" our predictors into regions, one for each class of the output.

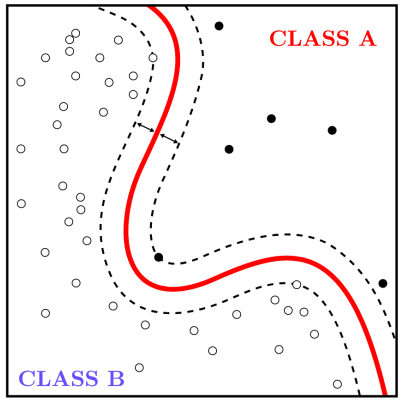

Figure 4: Visualisation of the support vector machine.

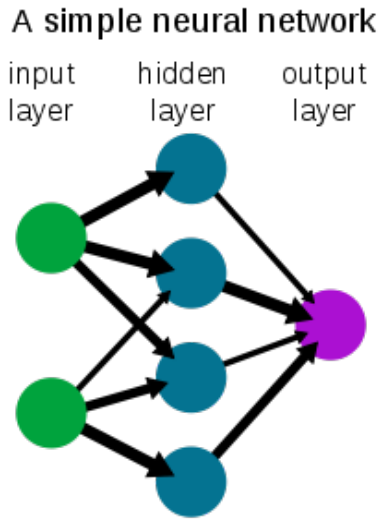

Figure 5: A simple neural network.

Results

The table summarises the accuracy between the classification tree, neural network, and support vector machine. From the table, we see that the support vector machine marginally outperforms neural network, and both of them outperform the classification tree.

Model	Accuracy
Classification Tree	0.8485096
Neural Network	0.8730094
Support Vector Machine	0.8961002

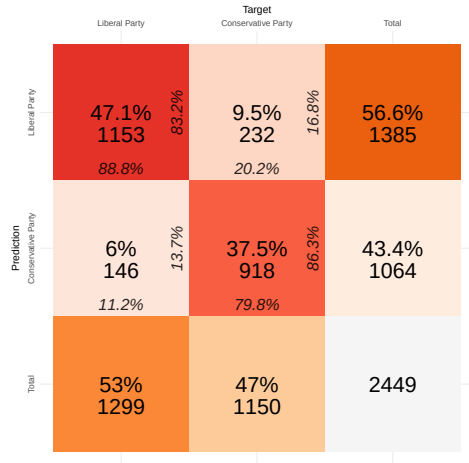

Figure 6: Confusion matrix for the classification tree.

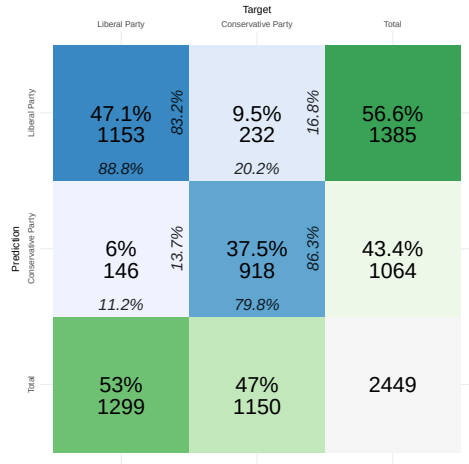

Figure 7: Confusion matrix for the support vector machine.

Conclusion

Our results showed that the support vector machine outperforms the other two models; however, all three models we tested scored fairly high in terms of accuracy. This shows that we can predict voter choice based on traits such as age, gender, and marital status.

Limitations

- For neural networks to perform well, and not overfit, we need enough data. Insufficient data (number of observations) was the biggest limitation. Hence, we could further improve the accuracy with more observations.
- The classification tree is sensitive to noise (a small change in the data can cause a large change in the tree structure), making it an unstable classifier.
- Similarly, the support vector machine does not perform well when there is noise, and complex structure in the data (unpredictable interdependence).

Summary of Report & Future Insights

1. Since the proportion of liberal party voters that support refugees is not 50%, the party should investigate to which direction the voters lean, to understand the stance that their voters are taking.
2. The medians are different with the women in the data set favoring Justin Trudeau more than men do. The party needs to investigate why, and work towards decreasing the divide between sentiment among the genders.
3. It is possible to predict which party a person will vote for based off of their traits, and the party needs to make sure they keep those that are already voting for them; however, the party should also reach out to the voters that could be swung over between parties.