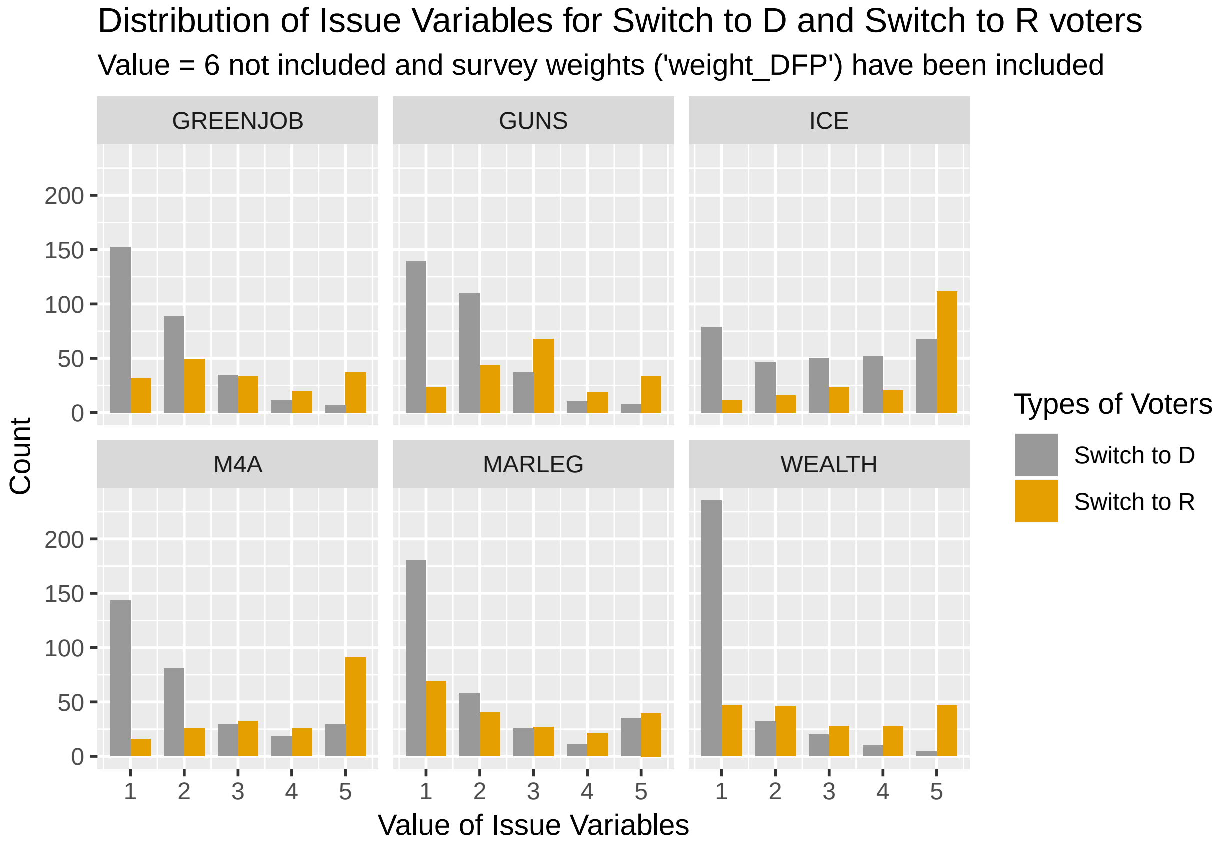
Mini Project 2

Abhilash Kuhikar, Dhruuv Agarwal, Darshan Shinde, Raj Thakkar

4 April 2019

Q1. How do Switch to D and Switch to R voters differ on the issue variables?

On which issue variables do Switch to D and Switch to R voters differ a lot? On which issue variables are they reasonably similar? Describe these differences.



For the analysis, value of 6 from each issue variable has been dropped as it means “Not sure” and thus doesn’t add any value in the exploratory analysis.

GREENJOB: A UN Environment program to preserve/restore environment quality.

For GREENJOB variable, the graph shows that most of the voters who Switch to D strongly supports that the point of giving every unemployed American who wants one a job building energy-efficient infrastructure. The count of voters decreases we move from strongly support (value = 1) to strongly oppose (value = 5).

On the other hand, the graph of Switch to R voters does not show any trend of their reaction. Hence, it is very hard to make any inference about their opinion on giving jobs in building energy efficient infrastructure to the unemployed Americans.

Guns: Gun control

For GUNS variable, the graph shows that most of the Switch to D voters believe that ‘It should be more difficult to buy all types of guns’. we can clearly see that count of voters decreases we move from strongly support (value = 1) to strongly oppose (value = 5).

On the other hand, the graph of Switch to R voters does not show any trend of their reaction. One important fact to note is that most number of Switch to R voters (value = 3) believe that the current regulations regarding the purchase of guns are about right.

ICE: Defunding Immingration and Customs Enforcement: Defund the ICE agency, which uses billions of tax payers money to lock up people with illegal migration status.

For ICE variable, the Switch to D voters have no clear opinion. In fact, almost equal number of voters support and oppose the defunding of Immigration and Customs Enforcement (ICE) agency. Hence, it is very hard to infer whether voters Switch to D either support or oppose the defunding of ICE agency.

In case of Switch to R voters, the number of voters increase as we move from strongly support the defunding of ICE agency (value = 1) to strongly oppose the defunding of ICE agency (value = 5)

M4A: Medicare for All

For M4A variable, Switch to D voters and Switch to R voters have opposite opinions. Most of the Switch to D voters support the Medicare for All program while most of the Switch to R voters oppose the Medicare for All program program.

MARLEG: Legalizing marijuana

From the graph for MARLEG variable, we can see that the trend is similar for both Switch to R and Switch to D voters although the magnitude of number of Switch to D voters is quite greater than number of Switch to R voters. Most of the voters (both Switch to R and Switch to D) support fully legalizing marijuana at the national level while least number of voters oppose fully legalizing marijuana at the national level

Wealth: Tax on wealth over $100 million

From the graph of WEALTH variable, the most obvious observation is Switch to D voters strongly support the WEALTH program which states 'there should a tax on wealth over $100 million'.

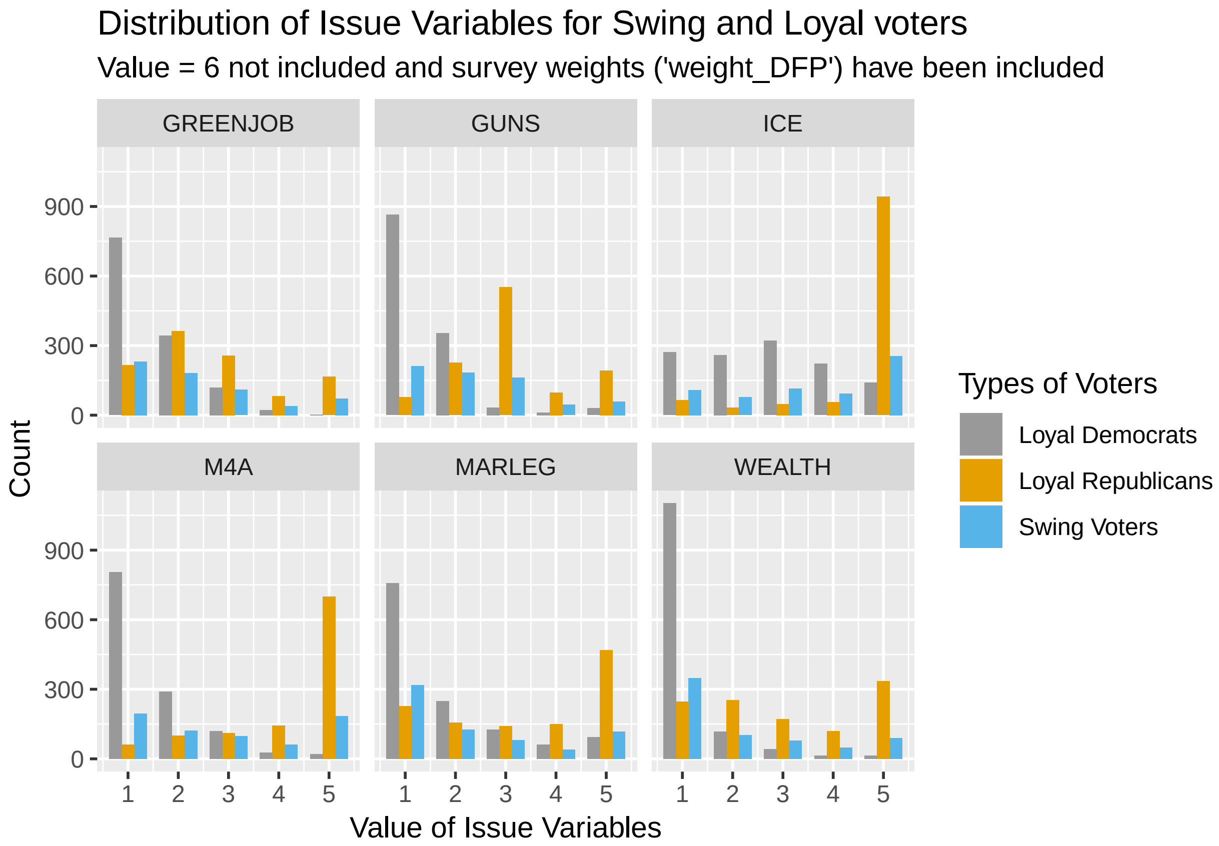
While on the other hand, Switch to R voters don't have any common opinion whether there should be a tax on wealth over 100 million or not.

To conclude, for GREENJOB, GUNS, WEALTH, ICE and M4A variables, Switch to D and Switch to R voters differ a lot. For MARLEG variable, the opinion exhibited by both Switch to D and Switch to R voters are reasonably similar with most of the voters supporting the corresponding program.

Out of the all 6 variables on which respondents were asked to express their view, ICE is the only variable where Switch do D voters do not have any clear opinion about the corresponding program. Other than ICE program they clearly support all the other programs corresponding to the issue variables.

While on the other hand, except ICE and M4A variables, there is no variable for which Switch to R voters have a common opinion about the corresponding program. In case of M4A program, opinions of Switch to D voters and Switch to R voters have opposite opinions.

Q2. How do swing voters differ from loyal Democrats and loyal Republicans on the issue variables?



From the graphs above, we can see that for the issue variables, in general Loyal Democrats are concentrated on the lower values (indicating strong support) while Loyal Republicans are concentrated on the higher values (generally strongly oppose).

Swing voters follow two hypothesis:

1 . Swing voters think more like Democrats for issue variables WEALTH, MARLEG, GUNS; and think more like Republicans for issue variables ICE and GREENJOB.

2 . Swing voters for issue variable M4A, are split with some proportion supporting both parties and thus are equally divided.

[exception: Loyal Republicans for GREENJOB, have slight support as distribution is more inclined towards low values(more support). So both parties are in support but Democrats are in much stronger support]

GREENJOB: A UN Environment program to preserve/restore environment quality.

We can see, the Loyal Democrats strongly support this cause, while the Loyal republicans show some support. This shows Democrats care about improving the environment.

swing voters some support the cause/policy, but as both parties do support the cause this was expected. But if we have to pick, we can say the distribution follows Republicans more as the peaks are not that skewed towards support as in Democrats.

Guns: Gun control

We can see, the Loyal Democrats strongly support this cause, while the Loyal Republicans don't have a strong view in general but small peaks indicate resistance against this control on guns.

The swing voters in general don't have a strong view again, but a small trend is there, which indicates more of total swing voters, strongly support this cause like the Democrats.

ICE: Defunding Immingration and Customs Enforcement: Defund the ICE agency, which uses billions of tax payers money to lock up people with illegal migration status.

As we can see from the plot, voters loyal to Republicans oppose this cause, while Loyal Democrats have in general no strong response to this policy/agency. This is in line with Republicans, strongly believe in stricter immigration policies and so oppose the reduction of funds to this cause.

The swing voters in general have no strong response but small peaks indicate swing voters follow Republicans more on this, and oppose the reduction of funds for stricter immigration policies.

M4A: Medicare for All

We can see, the Loyal Democrats strongly support this cause, while the Loyal Republicans are against this.

The swing voters in general don't have a strong view again, but the distribution looks split with portion of the swing voters population who follow the thought of the two parties.

MARLEG: Legalizing marijuana

We can see, the Loyal Democrats strongly support this cause, while the Loyal Republicans don't have a strong view in general but have some portion of voters against the legalization of marijuana.

The swing voters in general don't have a strong view again, but again two small peaks, bigger peak lies in the low region indicating that the swing voters generally share the Democrats thoughts and support the cause.

Wealth: Tax on wealth over $100 million

We can see, the Loyal Democrats strongly support this cause, while the Loyal Republicans don't have a strong view in general but have some portion of voters against this taxation.

The swing voters in general don't have a strong view again, but a small peak indicates strongly support this cause like the Democrats.

Q3. What predicts being a swing voter?

Model with Issue variables used as predictors

Building the first model which will only use issue variables as predictors. We will add predictors 1 at a time to see the effect of adding predictors to the model. There will be many models that we will need to build if we start from choosing 1 variable at a time from 6 issue variables to choosing 6 variables at a time from 6 issue variables. The total number of model will be:

$$no.of models = 6C1 + 6C2 + 6C3 + 6C4 + 6C5 + 6C6$$

$$no.of models = 6 + 15 + 20 + 15 + 6 + 1 = 63$$

Instead of creating 63 different models for comparison, we are making a safe assumption that the order of adding a variable in the model doesn't make a lot of difference. For simplicity, we will be adding variables in the model in the order of decreasing value of absolute value of correlation with Swing Voter variable.

From the ggpairs plot in the appendix, we can see that the order of variables in decreasing order of absolute value of correlation with SwingVoters variable is as below:

1. GREENJOB (Corr: 0.0844)
2. GUNS (Corr: 0.0559)
3. MARLEG (Corr: -0.0528)
4. M4A (Corr: 0.0337)
5. ICE (Corr: -0.0196)
6. WEALTH (Corr: 0.0193)

To understand how good the model with 1 predictor is we will do predictions on our training dataset and check the accuracy

```
## [1] "Train accuracy for model with 1 predictor is: 81.14 %"
```

To understand how good the model with 2 predictors is we will do predictions on our training dataset and check the accuracy

```
## [1] "Train accuracy for model with 2 predictors is: 81.65 %"
```

We can see that there is an improvement in accuracy (~0.5%) when 2 predictors are used in the model as compared to when only 1 predictor is used.

To understand how good the model with 3 predictors is we will do predictions on our training dataset and check the accuracy

```
## [1] "Train accuracy for model with 3 predictors is: 81.69 %"
```

We can see that there is a slight increase in accuracy (~0.04%) when 3 predictors are used in the model as compared to when 2 predictors are used.

To understand how good the model with 4 predictors is we will do predictions on our training dataset and check the accuracy

```
## [1] "Train accuracy for model with 4 predictors is: 81.88 %"
```

We can see that there is an improvement in accuracy (~0.2%) when 4 predictors are used in the model as compared to when 3 predictors are used.

To understand how good the model with 5 predictors is we will do predictions on our training dataset and check the accuracy

```
## [1] "Train accuracy for model with 5 predictors is: 81.82 %"
```

We can see that there is a slight decrease in accuracy (~0.06%) when 5 predictors are used in the model as compared to when 4 predictors are used.

To understand how good the model with 6 predictors is we will do predictions on our training dataset and check the accuracy

```
## [1] "Train accuracy for model with 6 predictors is: 81.95 %"
```

We can see that there is an improvement in accuracy (~0.13%) when 6 predictors are used in the model as compared to when 5 predictors are used.

For simplicity the results of accuracy (in %) for models with different number of issue variables used as predictors have been listed below in a table:

Table 1: Accuracy Table for Model with Issue variables

Model	Accuracy
Model with 1 predictor (GREENJOB)	81.14
Model with 2 predictors (GREENJOBS and GUNS)	81.65
Model with 3 predictors (GREENJOBS, GUNS and MARLEG)	81.69
Model with 4 predictors (GREENJOBS, GUNS, MARLEG and M4A)	81.88
Model with 5 predictors (GREENJOBS, GUNS, MARLEG, M4A and ICE)	81.82
Model with 6 predictors (GREENJOBS, GUNS, MARLEG, M4A, ICE and WEALTH)	81.95

Model with populism variables used as predictors

Building the second model which will only use populism variables as predictors. The populism variables also take values from 1 to 5 where 1means strongly agree and 5 means strongly disagree.

We will add predictors 1 at a time to see the effect of adding predictors to the model. There will be many models that we will need to build if we start from choosing 1 variable at a time from 3 populism variables to choosing 3 variables at a time from 3 populism variables. The total number of model will be:

$$no.ofmodels = 3C1 + 3C2 + 3C3$$

$$no.ofmodels = 3 + 3 + 1 = 7$$

Instead of creating 7 different models for comparison, we are making a safe assumption that the order of adding a variable in the model doesn't make a lot of difference. For simplicity, we will be adding variables in the model in the order of decreasing value of absolute value of correlation with SWing Voter variable.

From the ggpairs plot in the appendix, we can see that the order of variables in decreasing order of absolute value of correlation with SwingVoters variable is as below:

1. POP_1 ("It doesn't really matter who you vote for because the rich control both political parties.")(Corr: -0.133)
2. POP_2 ("The system is stacked against people like me.")(Corr: -0.0773)
3. POP_3 ("I'd rather put my trust in the wisdom of ordinary people than in the opinions of experts and intellectuals.")(Corr: -0.0182)

To understand how good the model with 1 predictor is we will do predictions on our training dataset and check the accuracy

```
## [1] "Train accuracy for model with 1 predictor is: 80.67 %"
```

To understand how good the model with 2 predictors is we will do predictions on our training dataset and check the accuracy

```
## [1] "Train accuracy for model with 2 predictors is: 80.83 %"
```

We can see that there is an improvement in accuracy (~0.16%) when 2 predictors are used in the model as compared to when only 1 predictor is used.

To understand how good the model with 2 predictors is we will do predictions on our training dataset and check the accuracy

```
## [1] "Train accuracy for model with 3 predictors is: 80.98 %"
```

We can see that there is an improvement in accuracy (~0.15%) when 3 predictors are used in the model as compared to when 2 predictors are used.

For simplicity the results of accuracy (in %) for models with different number of populism variables used as predictors have been listed below in a table:

Table 2: Accuracy Table for Model with Populism variables

Model	Accuracy
Model with 1 predictor (POP_1)	80.67
Model with 2 predictors (POP_1 and POP_2)	80.83
Model with 2 predictors (POP_1, POP_2 and POP_3)	80.98

From the 2 tables above, we can say that model with 6 issue variables is the best from the models in which different number of issue variables are used as predictors. For models with different number of populism variables, the model with all the 3 populism variables is the best. From 2 models, one with all the 6 issue variables and other with all the 3 populism variables, the model with all the 6 issue variables is better than the model with 3 populism variables as it has better train accuracy.

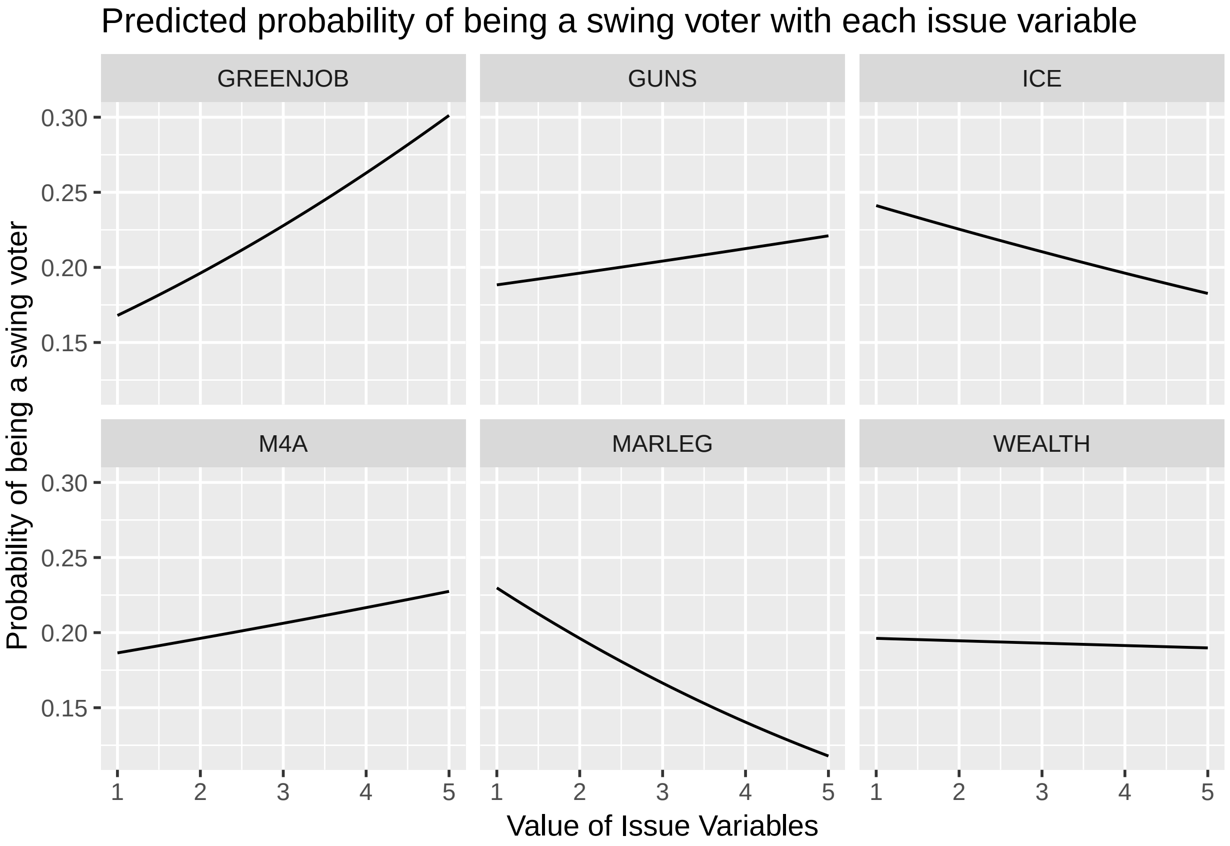
The 2 best models based on Issue variables and Populism variables are displayed below:

Best model with Issue Variables:

$$Pr(SwingVoters) = logit^{-1}(-1.25 + 0.19 * GREENJOB + 0.05 * GUNS - 0.2 * MARLEG + 0.06 * M4A$$
$$-0.09 * ICE - 0.01 * WEALTH)$$

Plotting the probability of being a swing voter based on Issue Variables

To plot the probability of being a swing voter with different values of issue variable, other issue variables are set to their median values. For example while predicting the probability of being a swing voter for different values of variable GREENJOB, other issue variables namely M4A, WEALTH, MARLEG, ICE and GUNS are set to their median values.



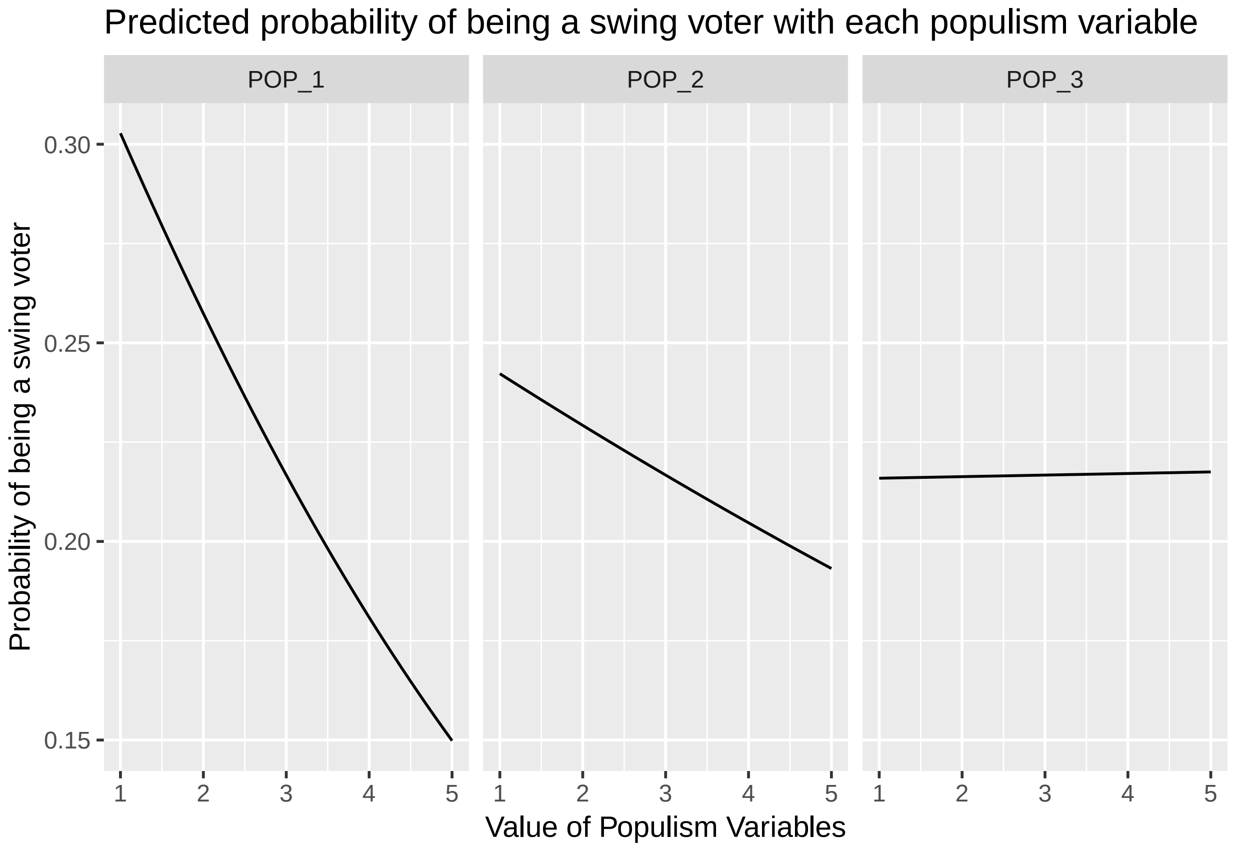
From the graph above we can see that the probability of being a swing voter changes to most with respect to different values of variable GREENJOB while the probability changes the least with respect to different values of variable WEALTH. This is inline with the magnitude of coefficients of the issue variables as shown in the equation above.

Best model with Populism Variables:

$$Pr(SwingVoters) = -0.40 - 0.23 * POP\_1 - 0.07 * POP\_2 + 0.002 * POP\_3$$

Plotting the probability of being a swing voter based on Populism Variables

To plot the probability of being a swing voter with different values of populism variable, other populism variables are set to their median values. For example while predicting the probability of being a swing voter for different values of variable POP_1, other populism variables namely POP_2, POP_3 are set to their median values.



From the graph above we can see that the probability of being a swing voter changes to most with respect to different values of variable POP_1 while the probability changes the least with respect to different values of variable POP_3. This is inline with the magnitude of coefficients of the populism variables as shown in the equation above.

Conclusion

The Switch to D and Switch to R voters differ on the issue variables GREENJOB, GUNS, ICE, M4A and WEALTH. While they are reasonably similar for the issue variable MARLEG.

Swing voters follow two hypothesis:

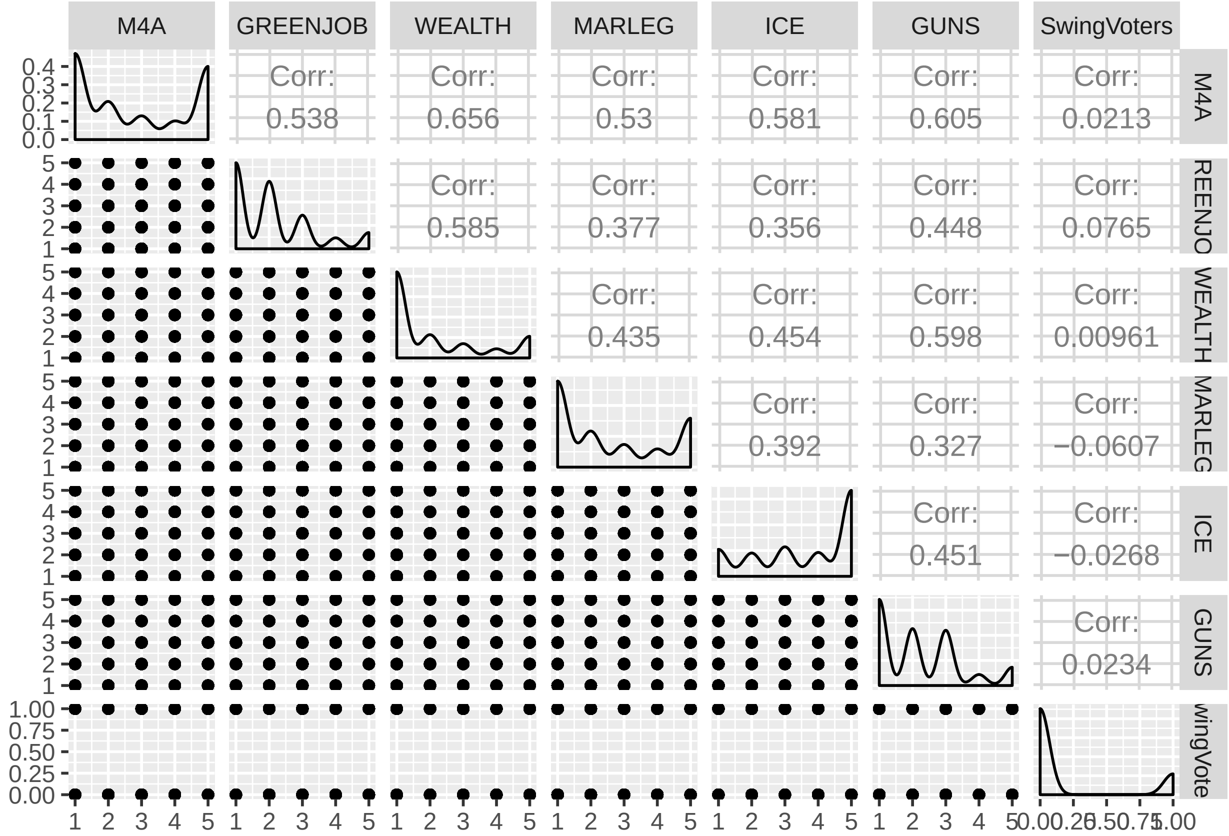
1 . Swing voters think more like Democrats for issue variables WEALTH, MARLEG, GUNS; and think more like Republicans for issue variables ICE and GREENJOB.

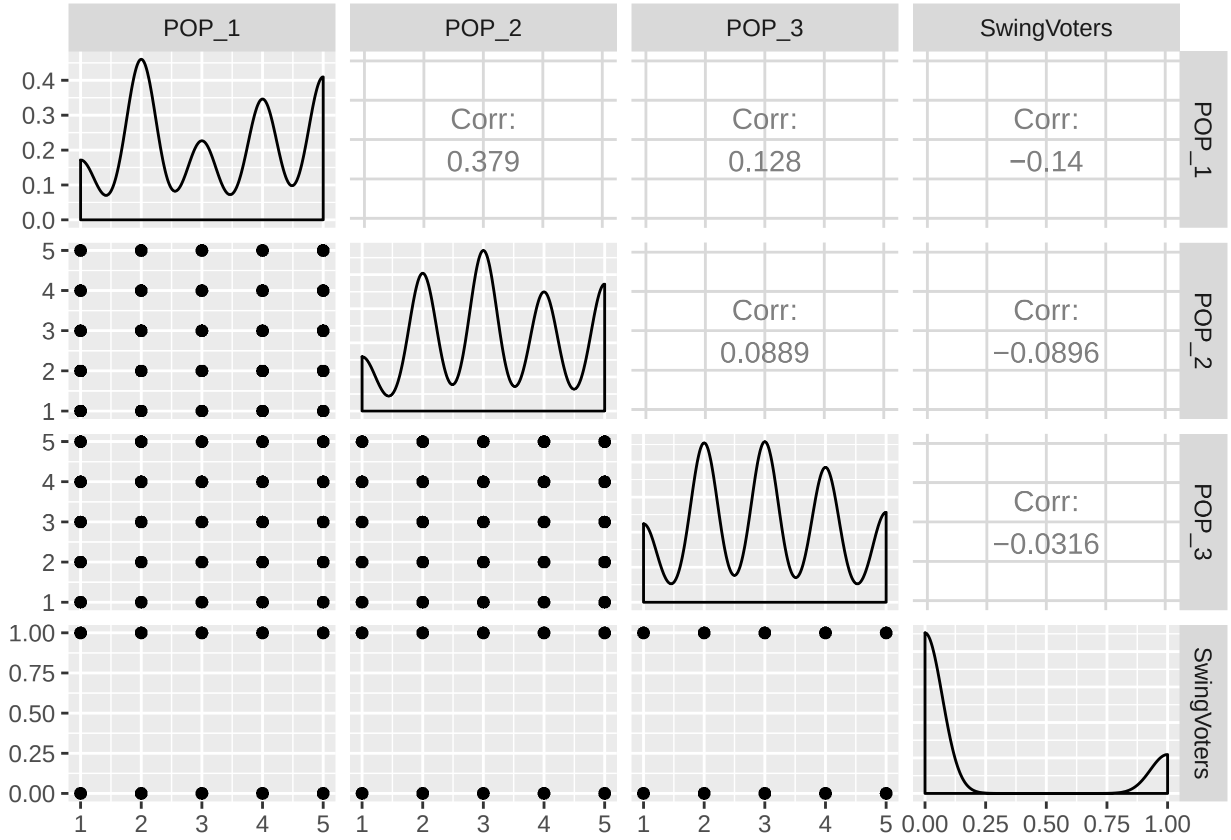
2 . Swing voters for issue variable M4A, are split with some proportion supporting both parties and thus are equally divided.

[exception: Loyal Republicans for GREENJOB, have slight support as distribution is more inclined towards low values(more support). So both parties are in support but Democrats are in much stronger support]

All the issue variables need to be used together to predict Swing voters for model with highest train accuracy. Similarly, all the populism variables need to be used together for the best model. The model with all the issue variables used together performs better than the model with all the populism variables used together. The highest training accuracy achieved is 81.95% approximately when all the 6 issue variables are used in the model.

Appendix


M4A
GREENJOB
WEALTH
MARLEG
ICE
GUNS
SwingVoters
Corr:
0.538
Corr:
0.656
Corr:
0.53
Corr:
0.581
Corr:
0.605
Corr:
0.0213
M4A
Corr:
0.585
Corr:
0.377
Corr:
0.356
Corr:
0.448
Corr:
0.0765
REENJO
Corr:
0.435
Corr:
0.454
Corr:
0.598
Corr:
0.00961
WEALTH
Corr:
0.392
Corr:
0.327
Corr:
−0.0607
MARLEG
Corr:
0.451
Corr:
−0.0268
ICE
Corr:
0.0234
GUNS
wingVote
0.4
0.3
0.2
0.1
0.0
5
4
3
2
1
1.00
0.75
0.50
0.25
0.00
1 2 3 4 5
0.00 0.25 0.50 0.75 1.00


POP_1
POP_2
POP_3
SwingVoters
Corr:
0.379
Corr:
0.128
Corr:
−0.14
Corr:
0.0889
Corr:
−0.0896
Corr:
−0.0316
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
0.00
0.25
0.50
0.75
1.00