7 Dwarves Final Project: Logistic Regression Model to Predict Voter Choice for 2020 Election

Team 7 Dwarves - Tali Nam, Malika Rawal, Morgan Bernstein, Lisa Zuo

Introduction and Data

Background and Research Question

Our group is extremely interested in politics, especially with the primary elections that are coming up right around the corner. We want to investigate why people voted the specific ways they did, focusing on their policy beliefs and non-political identities. One may assume that people vote based on a combination of policy positions and social identities. However, as Jenke and Huettel assert in their article, "Issues or Identity? Cognitive Foundations of Voter Choice," identity and policy positions actually compete to determine voter preference. We either vote based on our identity or policy positions, not both. If we follow their model of voter choice based on cognitive science, the increasing political polarization could be attributed to more people voting based on identity.

Can a statistical analysis of voter data could lead to additional insight into how important policy stances are in voter choice in comparison to identity-based metrics? We hypothesize that identity-based predictors, like whether or not a subject lives in a rural area, maybe better predictors of who someone voted for than policy-based predictors, like if a subject favors or opposes background checks for gun purchases.

The Data

Description

The source of the data set is ANES (American National Election Studies). This data is from the 2020 national election and used 8,280 pre-election surveys as well as 7,449 post-election surveys (ANES, 2020). The subjects were random U.S eligible voters and the data was collected through the internet, phone, and video. Our response variable is column V201033, which

corresponds to the survey question: For who does the respondent intend to vote for president? The possible answer choices are -9: Refused, -8: Don't Know, -1: Inapplicable, 1: Joe Biden, 2: Donald Trump, 3: Jo Jorgensen, 4: Howie Hawkins, 5: Other Candidate, 11: Specified as don't know, and 12: Specified as refused. We are planning to run a logistic regression, therefore we dropped all respondents who indicated intent to vote for someone other than Joe Biden or Donald Trump (or if they specified refused/don't know).

We removed 1505 observations from the initial 8280 to work with 6775 observations, so about 20% of the data was dropped, indicating about 20% of respondents who indicated intent to vote for a presidential candidate other than Joe Biden or Donald Trump.

Predictor Variables

We chose a subset of survey questions that we believed would be representative of people's political views and therefore provide good predictions for who they intended to vote for for president. We only decided to use these few questions as a subset of the original dataset, which includes over 1,700 questions/predictor variables. We felt that a smaller subset of the data would be easier to handle, easier to interpret and give a more straightforward prediction of a respondent's vote in the 2020 election. We wanted to make sure the survey questions we chose represented many facets of life both political ideologies and the identity of the respondent. Therefore, the 7 predictor variables we decided to include in the model are: Self-Placement on Abortion (V201336), Favor/Oppose Background Checks for Gun Purchases (V202339), Tightening Federal Budget Spending on Border Security (V201306), Self-Identified Race/Ethnicity (V201549x), Do You Currently Live in a Rural or Urban Area (V202355), Respondent Age (V201507x), What is Your Political Alignment (V202073).

Cleaning

After splitting the data into training (75%) and testing (25%), other steps in our data cleaning process included selecting only the specific predictors we wanted in the model along with our response variable, and mutating the column names so they are more representative of the survey question being asked, rather than the long survey question number. We also made sure to make all of the survey questions to factors (except for age) to represent them as categorical variables with multiple levels, rather than discrete numeric predictors. In addition, many of the survey questions gave 1 or more options for the respondent to not indicate their answer, such as a "don't know" or "refused" column. We believe these specific levels of the categorical variables would be better off put together into an "other" column, so we also included this specification in our data wrangling.
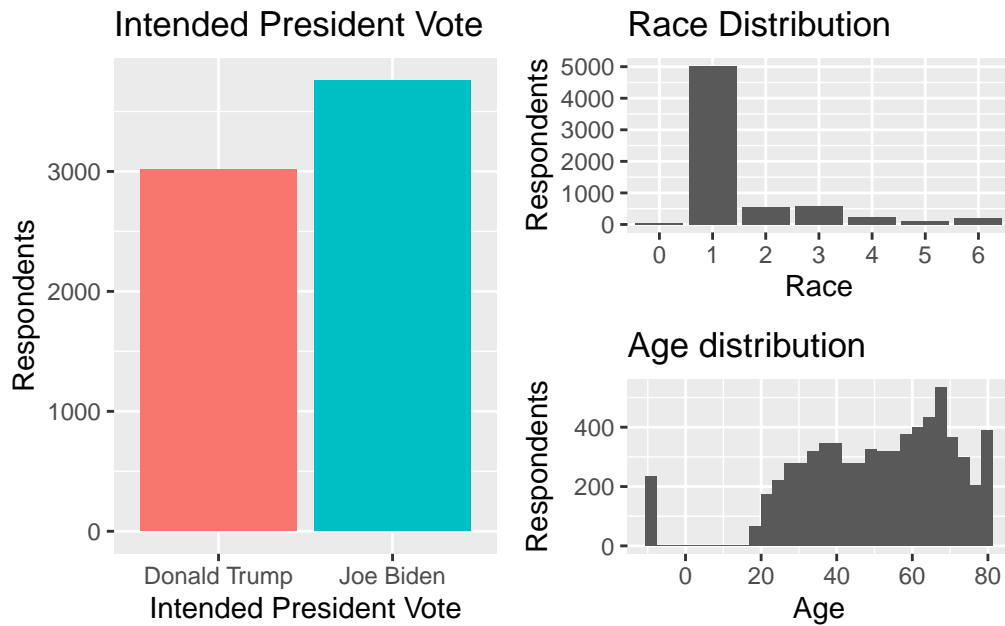
```
Rows: 6,775
Columns: 8
$ abortion            <dbl> 2, 4, 1, 3, 4, 3, 4, 1, 2, 4, 4, 3, 4, 4, 2, 4, 3~
$ gun_backgroundchecks <dbl> 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1~
```
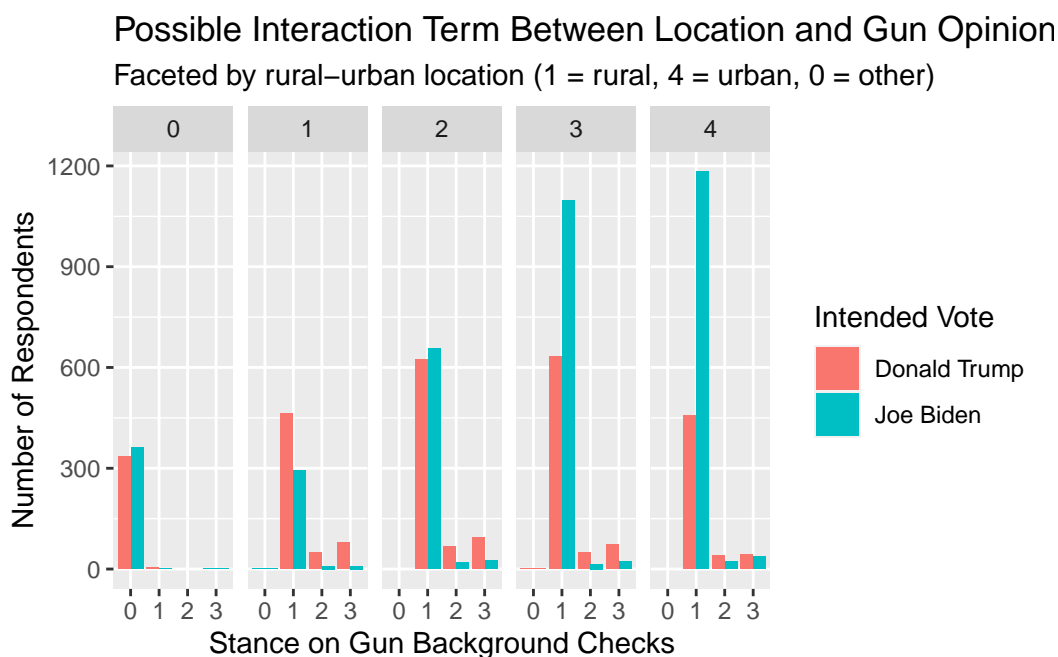
```
$ fedbudget          <dbl> 1, 2, 3, 1, 2, 1, 3, 3, 3, 1, 1, 1, 3, 3, 3, 1, 1~
$ rural_urban        <dbl> 4, 3, 2, 4, 3, 2, 2, 4, 2, 4, 2, 2, 3, 3, 3, 2, 1~
$ race               <dbl> 3, 1, 4, 5, 1, 1, 1, 3, 1, 1, 2, 3, 1, 1, 1, 3, 3~
$ age                <dbl> 46, 40, 41, 72, 71, 37, 45, 37, 55, 30, 41, 66, 5~
$ political_align    <dbl> 6, 2, 3, 5, 4, 4, 99, 99, 4, 6, 4, 6, 3, 2, 3, 6,~
$ president_vote     <dbl> 2, 1, 1, 2, 1, 2, 2, 1, 1, 2, 1, 2, 1, 1, 1, 2, 1~
```

Exploratory Data Analysis

The response variable (distribution of who respondents intend to vote for President) follows a binomial distribution, which is approximately 50-50. Slightly more respondents indicated their intent to vote for Joe Biden over Donald Trump. This indicates that respondents in the survey are generally representative of U.S. adults, as slightly more Americans did vote for Biden over Trump.

Some significant discoveries from our exploratory analysis of individual variables in our data set were in the distribution of race and age legislation opinions. First, the distribution of race, it was highly skewed in favor of white (Response 1). Additionally, the distribution of age shows that it is approximately evenly distributed, but there are fewer young (under the age of 25) Americans represented. This indicates that there may be some bias in the way the survey was collected. We can still make conclusions from this dataset; however, we wanted to highlight where there is less data which might lead to less accurate predictions for these groups.

Possible Interaction Term Between Location and Gun Opinion
Faceted by rural–urban location (1 = rural, 4 = urban, 0 = other)

We explored a possible interaction term between rural_urban location and stance on background checks for guns. When respondents indicated that they resided in a rural location (1), they were more likely to intend to vote for Donald Trump, even if they held opinion 1 (favoring requiring background checks for gun purchases at gun shows or other private sales) as compared to respondents from other locations.

Methodology

Checking Conditions and Choosing the Type of Model

We are running a logistic regression model to predict a respondent's response to the question: For whom does the respondent intend to vote for president? As mentioned, we are only looking at respondents who intended to vote for Joe Biden (1) or Donald Trump (2). The three conditions for logistic regression are linearity, randomness, and independence. Given that the survey was conducted of random U.S eligible voters, we know from the context of the analysis that the randomness condition is met. This also gives us insight into the independence condition, since the surveys were conducted randomly it is a reasonable assumption that one person's survey responses did not affect another person's responses, therefore we can also conclude that the independence condition is met.
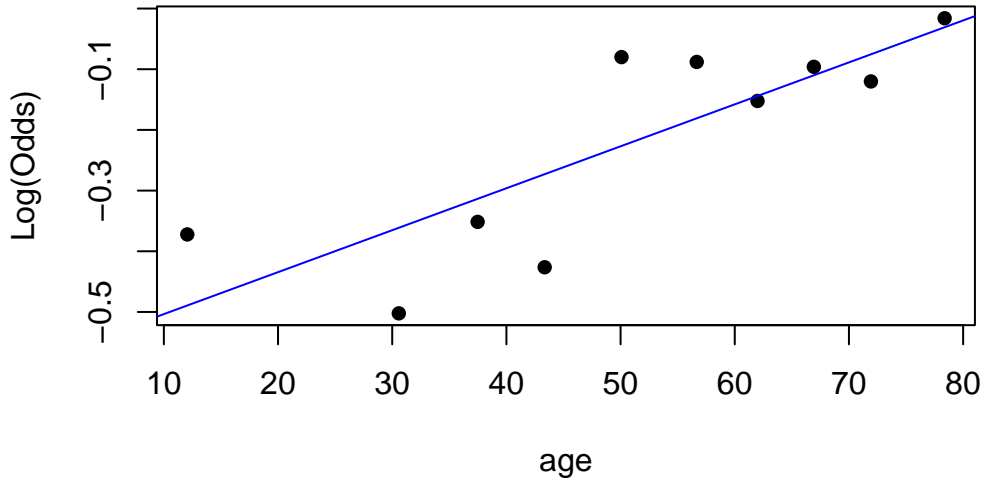
The only numeric predictor in our model is age, so this is the one we calculated the empirical logit for. All other variables are categorical which we made into factors, therefore we will not assess their linearity. The empirical logit graph for age shows no obvious deviations from a linear relationship with the response, therefore we will assume the condition of linearity is met.

Interactions

There were interaction terms that we considered including in our model, including between rural-urban identity and stance on gun-background checks and age and political alignment. Additionally, we knew that in more rural areas, hunting is usually more popular than in urban areas, leading to increased gun access. Our EDA between rural-urban identity and gun background checks also indicates the possibility of an interaction term. Lastly, we read an article stating as you grow older you tend to be more conservative due to new trends in society, and fear of economic stability (401(k)), amongst other rationales. All of these thought processes helped us determine the interaction terms used.

However, all models that included interaction effects produced higher AIC and BIC values than our original model which did not have any interactions. AIC and BIC are both measures of model fit where lower values are considered favorable and represent a better model fit to the data. Therefore, because of the lower AIC and BIC values for our model without interactions, we did not include any of the interaction terms in our final model. The model without interactions also had a lower area under the ROC curve, which is another measurement

of model fit for logistic regression where a higher value is preferred. See the appendix for more details.

Selecting our Model

We ultimately chose the model with the lowest BIC/AIC value which was the model with no interaction terms. This model uses the predictor's Self Placement on Abortion, Favor/Oppose Background Checks for Gun Purchases, Tightening Federal Budget Spending on Border Security, Self-Identified Race/Ethnicity, Rural or Urban identity, Age, and Self-Identified Political Alignment to predict whether a respondent planned to vote for Trump or Biden in the 2020 election.

term	estimate	std.error	statistic	p.value
(Intercept)	0.7334315	1.3432609	0.5460083	0.5850603
age	-0.0129114	0.0025594	-5.0447398	0.0000005
political_align_X.8	-10.0858437	324.7448182	-0.0310578	0.9752235
political_align_X1	-1.8452883	0.9009858	-2.0480769	0.0405525
political_align_X2	-2.3336210	0.8547998	-2.7300205	0.0063330
political_align_X3	-2.0223896	0.8492470	-2.3813915	0.0172474
political_align_X4	-0.7518008	0.8356909	-0.8996160	0.3683247
political_align_X5	0.5724233	0.8408844	0.6807396	0.4960363
political_align_X6	2.0555138	0.8464895	2.4282802	0.0151706
political_align_X7	2.1705925	0.9220301	2.3541450	0.0185654
political_align_X99	-0.1539753	0.8377137	-0.1838041	0.8541671
abortion_X1	0.6164121	0.2810230	2.1934576	0.0282744
abortion_X2	0.3890293	0.2486384	1.5646386	0.1176677
abortion_X3	-0.1622702	0.2548773	-0.6366599	0.5243464
abortion_X4	-0.8817425	0.2380450	-3.7041000	0.0002121
gun_backgroundchecks_X1	-0.2603096	1.1274367	-0.2308862	0.8174032
gun_backgroundchecks_X2	0.6088799	1.1610647	0.5244151	0.5999898
gun_backgroundchecks_X3	0.7950461	1.1467377	0.6933112	0.4881143
fedbudget_X1	0.9865275	0.9828092	1.0037833	0.3154831
fedbudget_X2	-1.9845241	0.9917453	-2.0010422	0.0453878
fedbudget_X3	-0.6340558	0.9832417	-0.6448626	0.5190162
rural_urban_X1	0.6164699	1.1289252	0.5460679	0.5850192
rural_urban_X2	0.0287435	1.1263824	0.0255184	0.9796415
rural_urban_X3	-0.0977908	1.1250408	-0.0869220	0.9307335
rural_urban_X4	-0.1597203	1.1255474	-0.1419045	0.8871554
race_X1	0.7853995	0.5118131	1.5345435	0.1248960
race_X2	-2.2413486	0.5545028	-4.0420871	0.0000530
race_X3	-0.3473468	0.5301498	-0.6551863	0.5123478
race_X4	0.0949158	0.5621074	0.1688570	0.8659091

term	estimate	std.error	statistic	p.value
race_X5	0.5767729	0.5887723	0.9796197	0.3272739
race_X6	0.2504564	0.5666611	0.4419862	0.6584992

Results

Model Fit and Predictive Power Assessment

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.952
```
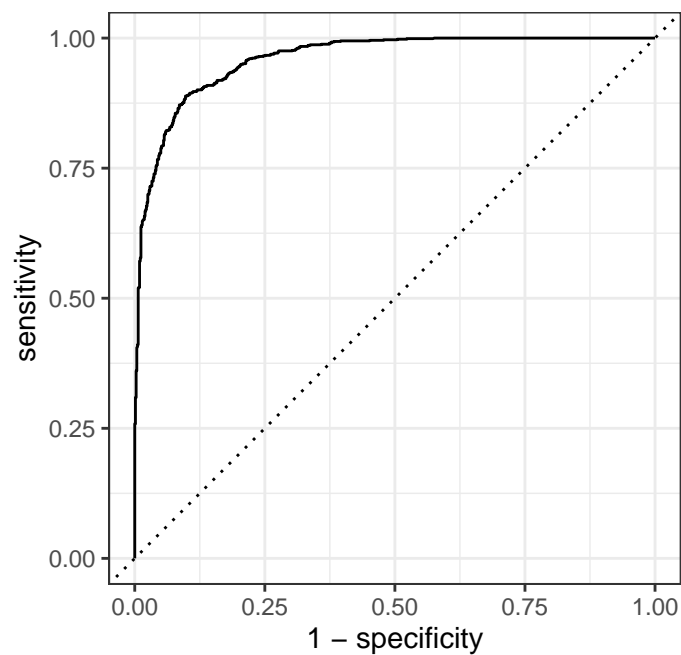


```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.960
```

When fit to the testing data, our model performs extremely well! It has an area under the ROC curve of .96, which corresponds to a very good model fit, as the AUC is a numerical

representation of how well our model classifies each observation into a binary output. Additionally, our model has a slightly higher AUC (.9604) when fit to the testing data compared with the AUC when fit to the training data (.9521). Both numbers show a good model fit, however since the model performs better on new data it hasn't seen, we have evidence that it is NOT overfitting to the training data.

We also evaluated the AIC and BIC values for our model. These were minimized in the model we decided to use compared with all other models tried, since AIC and BIC are measures of model fit that we want to minimize (see Appendix for details).

Age: For each additional year increase in age, on average, the odds of voting for Trump are expected to multiply by 0.987, holding all else constant.

Political Alignment Summary: When looking at the respondent's political alignment, we noticed that there is a trend from Slightly Liberal to Extremely Liberal with a dip in the middle for just Liberal. As a respondent becomes increasingly liberal only from slightly to extremely, the odds of voting for Trump are expected to decrease. For example, if the respondent's political alignment is "Slightly Liberal" rather than "Refused," the odds of voting for Trump are expected to multiply by 0.132, holding all else constant. In the case of "Liberal" rather than "Refused," the odds of voting for Trump are multiplied by only 0.097. However, in the case of "Extremely Liberal" rather than "Refused," the odds of voting for Trump are multiplied by 0.160, holding all else constant. (0.160 > 0.132 for Extremely in comparison to Slightly Liberal).

In the same way, the more conservative a respondent becomes, the odds of voting for Trump increase. If the respondent's political alignment is "Conservative" rather than "Refused", the odds of voting for Trump are expected to multiply by 7.810, holding all else constant. If the respondent's political alignment is "Extremely Conservative" rather than "Refused", the odds of voting for Trump multiply by 8.758, holding all else constant. (8.758 > 7.810 for Extremely Conservative in comparison to Conservative)

Abortion Summary: If the respondent's abortion beliefs are "By law, abortion should never be permitted" rather than "Other", then the odds of voting for Trump are expected to multiply by 1.85, holding all else constant. If the respondent's abortion beliefs are "By law a woman should always be able to obtain an abortion as a matter of personal choice" rather than "Other", the odds of voting for Trump are expected to multiply by 0.414, holding all else constant.

Federal Budget: If the respondent's federal budget beliefs are "Decreased" when answering "Should federal spending on tightening border security to prevent illegal immigration be increased, decreased, or kept the same" rather than "Other", the odds of voting for Trump are expected to multiply by 0.138, holding all else constant.

Race: If the respondent's race is "Black, non-Hispanic" rather than "Other", the odds of voting for Trump are expected to multiply by 0.106, holding all else constant.

```
# A tibble: 4 x 3
  president_vote .pred_class      n
  <fct>          <fct>        <int>
1 1              1              842
2 1              2               93
3 2              1               98
4 2              2              661
```

(President Vote 1 is Biden and 2 is Trump) The sensitivity of our model is .8709, and the specificity of our model is .9005. Since we have similar values for both, this means that our model has a fairly similar false negative and false positive rate, meaning it correctly (and incorrectly) predicts respondents to vote for both presidential candidates at a fairly similar rate. This is preferred in our model since we don't have a reason to minimize false positives or false negatives over the other. However, specificity is a bit higher than sensitivity in our case, therefore our model can correctly classify respondents as not voting for Donald Trump when they truly did vote for Joe Biden at a slightly higher rate.

Discussion and Conclusion

Most of our findings were in line with the political parties' opinions on polarizing issues in our country. When a respondent becomes increasingly liberal, their odds of voting for Donald Trump decrease from slightly to extremely liberal (with a dip below slightly for just Liberal). In the same way, the more conservative a respondent becomes, the more likely they are to vote for Trump. If a respondent's views on abortion are "By law, abortion should never be permitted," rather than "Other", then their odds of voting for Trump increase. However, with a response of "By law, a woman should always be able to obtain an abortion as a matter of personal choice" rather than "Other", the odds of voting for Trump are expected to decrease. This is an indication that it is possible that the more stringent you are about abortion policy, the more likely you are to have voted for Trump. There is also an indication that if a respondent desires less border security they are less likely to have voted for Trump. Lastly, we found that respondents who were "Black, non-Hispanic" are less likely to vote for Trump.

A main goal of our interest in this research project was to determine which factors of a person's identity or political alignment were the most influential in predicting which presidential candidate they voted for. Using a significance level of .05, we determined that the specific categorical predictors of age, political alignment, abortion, federal budget spending on border security, and race were significantly significant when predicting who a subject intended to vote for. Therefore, based on our predictors, we conclude that both policy and identify factors can be used to predict who a person will vote for. We would encourage campaigns to use this information to identify where they should campaign and what they should focus on. Given that federal budget spending on border security and abortion legislation are statistically significant predictors, political campaigns should highlight these aspects of their policy.

9

Ideas for Future Work: We believe that in the future, a more comprehensive way to determine how identity-based and policy-based predictors affect how a respondent votes would be to use smaller elections such as local ones as well as more analysis over time (such as using more than one year of data). In this case, the data was over the entire United States, so there were a lot of other factors to consider such as region. Additionally, adding more predictors to the model would increase the comprehensiveness of the report.

Potential Limitations: There are many potential limitations to the findings of this research project such as Undercoverage and Response Bias. There may be undercoverage due to the polls being taken via the internet, phone, and video but not on paper for those who do not have access. There may be response bias since all of the survey questions have "refused" and "don't know" creating a possibility for those with more extreme views to respond in the categories we filtered for. Additionally, the response variable determines who the respondent intended to vote for, not actually voted for. There could have been changes between the time of this survey to the actual election. Lastly, we only chose a few predictors from the wide variety of survey questions. A more comprehensive report would add additional predictors.

References:

Jenke, L., & Huettel, S. A. (2016). Issues or Identity? Cognitive Foundations of Voter Choice. Trends in cognitive sciences, 20(11), 794. https://doi.org/10.1016/j.tics.2016.08.013

American National Election Studies. 2021. ANES 2020 Time Series Study Full Release [dataset and documentation]. July 19, 2021 version.www.electionstudies.org

See data dictionary

Appendix

Model 2

term	estimate	std.error	statistic	p.value
(Intercept)	0.6612565	1.3559942	0.4876544	0.6257947
age	-0.0129294	0.0025646	-5.0414959	0.0000005
gun_backgroundchecks1__x__rural__urban1	0.4104559	1.2552528	0.3269906	0.7436750
gun_backgroundchecks2__x__rural__urban1	0.0660728	0.5905466	0.1118842	0.9109152
gun_backgroundchecks3__x__rural__urban1	12.1579098	535.4115253	0.0227076	0.9818835
gun_backgroundchecks1__x__rural__urban2	-0.1894178	1.2527548	-0.1512010	0.8798172
gun_backgroundchecks2__x__rural__urban2	0.8674007	0.4926820	1.7605691	0.0783114
gun_backgroundchecks3__x__rural__urban2	10.8344637	535.4113754	0.0202358	0.9838553
gun_backgroundchecks1__x__rural__urban3	-11.2481472	535.4126722	-0.0210084	0.9832390
gun_backgroundchecks2__x__rural__urban3	-10.0162463	535.4115375	-0.0187076	0.9850744
gun_backgroundchecks3__x__rural__urban3	-0.2804321	757.1859654	-0.0003704	0.9997045
gun_backgroundchecks1__x__rural__urban4	-0.4030394	1.2519721	-0.3219236	0.7475106
gun_backgroundchecks2__x__rural__urban4	0.6405873	0.5280362	1.2131505	0.2250723
gun_backgroundchecks3__x__rural__urban4	11.0382695	535.4114110	0.0206164	0.9835516
political_align_X.8	-11.0744099	535.4118523	-0.0206839	0.9834978
political_align_X1	-1.8696936	0.9028423	-2.0708972	0.0383684
political_align_X2	-2.3525887	0.8562003	-2.7477082	0.0060013
political_align_X3	-2.0392070	0.8504631	-2.3977608	0.0164956
political_align_X4	-0.7612272	0.8368999	-0.9095799	0.3630441
political_align_X5	0.5651510	0.8420501	0.6711608	0.5021181
political_align_X6	2.0409119	0.8476967	2.4075968	0.0160579
political_align_X7	2.1741099	0.9236445	2.3538383	0.0185807
political_align_X99	-0.1616807	0.8388381	-0.1927436	0.8471597
abortion_X1	0.6174498	0.2815051	2.1933874	0.0282795
abortion_X2	0.4015057	0.2492778	1.6106758	0.1072504
abortion_X3	-0.1612643	0.2557571	-0.6305371	0.5283432
abortion_X4	-0.8764080	0.2386228	-3.6727761	0.0002399
gun_backgroundchecks__X1	-0.0358423	1.2543347	-0.0285748	0.9772037
gun_backgroundchecks__X2	NA	NA	NA	NA
gun_backgroundchecks__X3	-10.2142932	535.4112807	-0.0190775	0.9847793
fedbudget_X1	1.0292848	0.9991344	1.0301765	0.3029272
fedbudget_X2	-1.9678444	1.0082269	-1.9517873	0.0509635
fedbudget_X3	-0.5905130	0.9995439	-0.5907824	0.5546662
rural__urban_X1	NA	NA	NA	NA
rural__urban_X2	NA	NA	NA	NA
rural__urban_X3	10.9270726	535.4112198	0.0204087	0.9837173
rural__urban_X4	NA	NA	NA	NA

| term	estimate	std.error	statistic	p.value
race__X1	0.8240216	0.5130271	1.6061950	0.1082311
race__X2	-2.1931312	0.5561343	-3.9435277	0.0000803
race__X3	-0.3130898	0.5313728	-0.5892093	0.5557209
race__X4	0.1167860	0.5640733	0.2070405	0.8359782
race__X5	0.6390320	0.5904677	1.0822472	0.2791427
race__X6	0.2607239	0.5677752	0.4592027	0.6460886
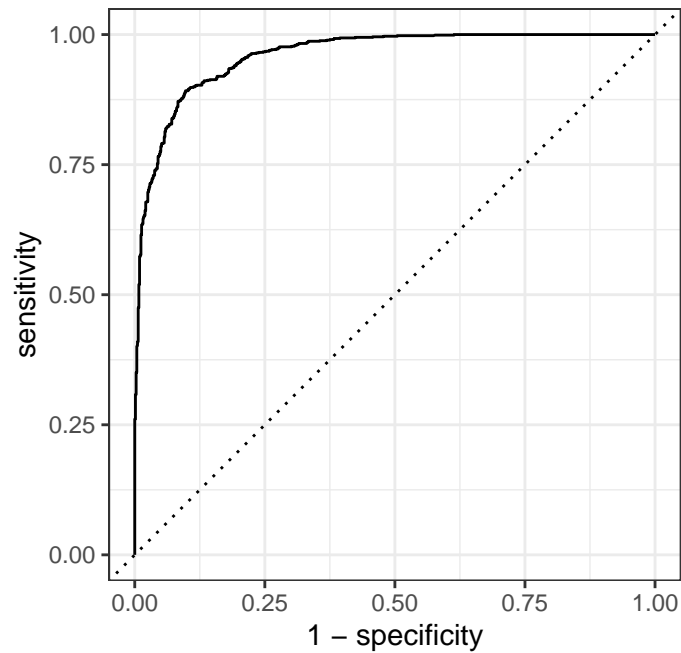

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.960
```

Model 3

term	estimate	std.error	statistic	p.value
(Intercept)	1.8707981	2.3268050	0.8040201	0.4213853
age	-0.0366726	0.0348734	-1.0515912	0.2929871
age_x_political_align-8	-0.1647362	5.4124146	-0.0304367	0.9757187
age__x_political_align1	0.0369037	0.0385315	0.9577535	0.3381871

term	estimate	std.error	statistic	p.value
age__x__political__align2	0.0229204	0.0363458	0.6306201	0.5282890
age__x__political__align3	0.0261547	0.0359705	0.7271152	0.4671554
age__x__political__align4	0.0180632	0.0351065	0.5145261	0.6068842
age__x__political__align5	0.0374464	0.0353716	1.0586549	0.2897570
age__x__political__align6	0.0206243	0.0359194	0.5741842	0.5658432
age__x__political__align7	0.0336631	0.0389421	0.8644399	0.3873463
age__x__political__align99	0.0207584	0.0352270	0.5892770	0.5556755
political__align__X.8	NA	NA	NA	NA
political__align__X1	-3.6795676	2.2446577	-1.6392556	0.1011600
political__align__X2	-3.5240810	2.1559620	-1.6345747	0.1021382
political__align__X3	-3.3768375	2.1392417	-1.5785208	0.1144460
political__align__X4	-1.7031768	2.0940774	-0.8133304	0.4160286
political__align__X5	-1.3419626	2.1074098	-0.6367829	0.5242662
political__align__X6	0.9952078	2.1443732	0.4641020	0.6425747
political__align__X7	0.4021472	2.3030900	0.1746120	0.8613845
political__align__X99	-1.2329466	2.1016762	-0.5866492	0.5574394
abortion__X1	0.6385144	0.2822521	2.2622130	0.0236842
abortion__X2	0.3923340	0.2494925	1.5725283	0.1158281
abortion__X3	-0.1626630	0.2558492	-0.6357768	0.5249219
abortion__X4	-0.8790514	0.2389708	-3.6784882	0.0002346
gun__backgroundchecks__X1	-0.3320989	1.1478307	-0.2893274	0.7723308
gun__backgroundchecks__X2	0.5309094	1.1798097	0.4499958	0.6527134
gun__backgroundchecks__X3	0.7149576	1.1668490	0.6127251	0.5400581
fedbudget__X1	0.9935260	0.9859151	1.0077196	0.3135891
fedbudget__X2	-1.9859482	0.9949546	-1.9960190	0.0459319
fedbudget__X3	-0.6411293	0.9864233	-0.6499535	0.5157222
rural__urban__X1	0.6855606	1.1495243	0.5963864	0.5509171
rural__urban__X2	0.0950036	1.1467527	0.0828458	0.9339742
rural__urban__X3	-0.0424642	1.1455455	-0.0370690	0.9704300
rural__urban__X4	-0.1005922	1.1459921	-0.0877774	0.9300536
race__X1	0.8872298	0.5174585	1.7145911	0.0864202
race__X2	-2.1427761	0.5598241	-3.8275882	0.0001294
race__X3	-0.2435065	0.5355804	-0.4546591	0.6493545
race__X4	0.2122481	0.5678332	0.3737860	0.7085635
race__X5	0.6702514	0.5938264	1.1286992	0.2590247
race__X6	0.3571206	0.5715655	0.6248114	0.5320949
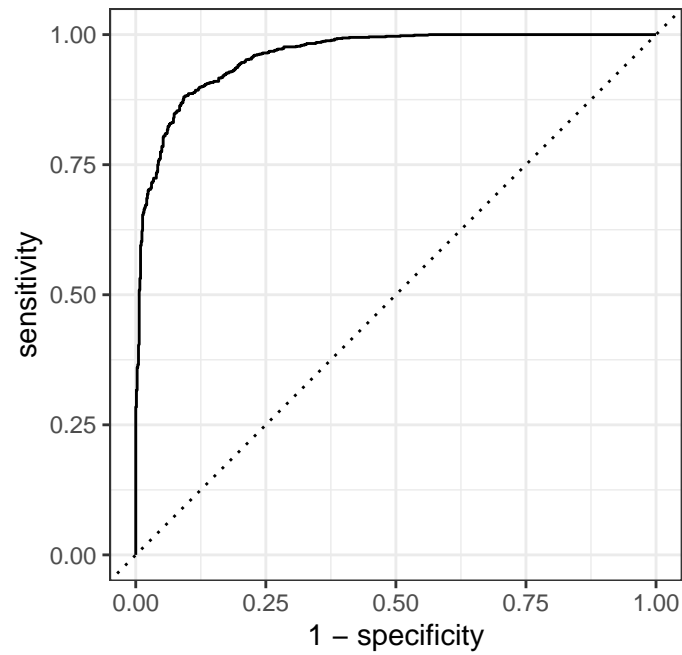
```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.960
```

AIC and BIC values for all models

Model 1:

```
# A tibble: 1 x 2
    AIC   BIC
  <dbl> <dbl>
1 2892. 3095.
```

Model 2:

```
# A tibble: 1 x 2
    AIC   BIC
  <dbl> <dbl>
1 2902. 3156.
```

Model 3:

```
# A tibble: 1 x 2
    AIC   BIC
  <dbl> <dbl>
1 2899. 3154.
```