Who Donates? Using Machine Learning to Predict Federal Donation Behavior

Nick Pangakis Xinyi Wang Ludwig Zhao

Contents

Executive Summary

Campaign contributions are an important facet of political behavior that affects who gets elected and what types of policy are implemented. When considered as investments and attempts to influence political outcomes, campaign contributions are a critical component of politics. Moreover, campaign contributions are a multi-billion-dollar political industry. From 2000 to 2020, for example, individual donors living in Pennsylvania contributed more than 23 billion dollars to campaigns for federal office. Non-federal contributions and donations by corporations only increase the significance of money in politics.

Is it possible to predict who will donate money to a political campaign? While social scientists have invested notable time and resources into understanding how to mobilize voters and how to change political attitudes (e.g., Broockman and Kalla 2016; Gerber and Green 2000), significantly less scholarship has investigated optimal strategies for identifying prospective voters and donors. The primary goal of this project is to use machine learning to predict federal campaign contribution behavior in the United States. First, we implement LASSO logistic regression as a baseline model. Second, we implement random forest. Across both methods, the target variable is a binary indicator for whether an individual donated money to a federal political campaign in 2020. The data for our analysis merge federal campaign contribution data, Pennsylvania voting records, governmental socio-demographic county data, and 2016 U.S. presidential election results.

Our findings indicate that LASSO logistic regression and random forest are very effective at predicting donation behavior. LASSO logistic regression correctly classifies 82.7 percent of test cases and random forest correctly classifies 92.8 percent of test cases. Both of these accuracy scores are notably higher than the 74.1 percent no information rate. Nevertheless, LASSO logistic regression is only able to correctly classify 61.9 percent of positive cases. Remarkably, however, random forest correctly classifies 99.9 percent of positive classes, which indicates that it is the superior model by far. As a result, we conclude that random forest is an extremely effective method for predicting campaign behavior. In our conclusion, we discuss implications of our analysis and next steps for the project.

1 Data Preparation

This study merges data from four sources:

1. First, we utilize data from the 2021 Pennsylvania voter file, which contains comprehensive administrative data on every individual registered to vote (i.e., over 8 million individuals) in the state of Pennsylvania in 2021. The data includes information on all voters including name, sex, date of birth, date of registration, voter status (i.e., active or inactive), registered political party, residential address, and complete voting history since 2012. We also impute the race of each voter using each voter's name and address.

2. Second, we access all individual-level campaign contributions to federal elections (2000-2020) from individuals living in Pennsylvania. This data comes from the Federal Election Commission (FEC), which is available for public download. Founded in 1975, the FEC is a federal organization created for the primary purpose of enforcing campaign finance law. In the name of campaign finance transparency, this federal agency has maintained standardized collections of all individual-level contributions for over twenty years. The raw data includes the name of the contributor, the contributor's zip code, the contributor's employer, the contributor's occupation, the date of the contribution, the donation amount, and which candidate or political organization received the donation. The raw donation data is hundreds of millions of observations for the whole country. For example, from 2017 to 2020 alone, there were 153,592,950 donations made to federal campaigns across the country.

3. Third, we draw on socio-demographic data from the American Community Survey (ACS) 2015-2019 (5-year estimates).. The ACS is an ongoing governmental survey that provides important information on a yearly basis about the United States. The survey results help determine how more than $675 billion in federal and state funds are distributed each year. For this analysis, we downloaded county-level data on total population, median income, percent white, percent Black, percent Hispanic, percent noncitizen, and percent college educated.

4. Finally, we use county-level election results from the 2016 U.S. Presidential Election. The data come from the MIT Election Lab.

1.1 Data Preprocessing

The full data cleaning procedures and code can be found in the .Rmd file located on GitHub. On GitHub, we include the finalized clean data. Please note that the raw data for this project is not included on GitHub because the raw data is hundreds of millions of observations stored in over 40 separate files. That being said, the data is all publicly available for download in order to replicate this study from raw data. The FEC data can be found here. The ACS data can be found here. The election results can be found here. The 2021 Pennsylvania voter file can be found here.

Brief comments on the data cleaning process:

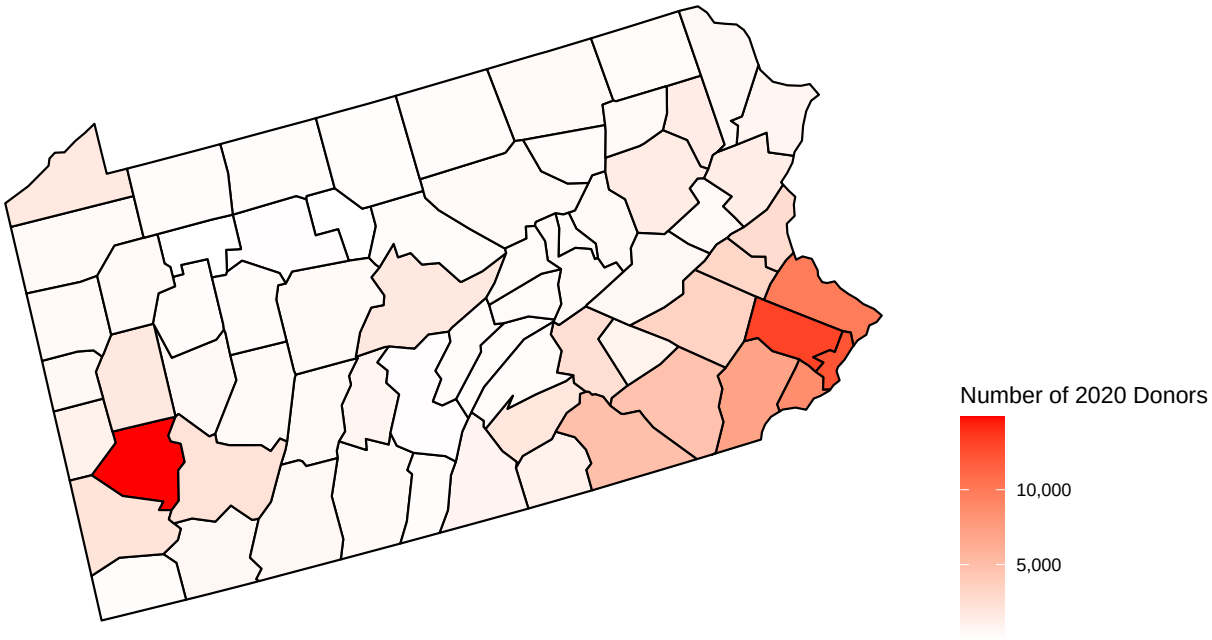
1. The final data is a merge of FEC campaign contribution data, the PA voter file, ACS socio-demographic county data, and 2016 U.S. presidential election results. In $Table1$ and $Table2$ in the Appendix, we display each numeric and categorical variable and provide a brief description of each variable.
2. After loading in the raw data for each year of FEC contributions, we group each donation by first name, last name, and zip code and then aggregate the total donation amount per year. After joining the donation data with the PA voter file, the resulting dataframe is a unique row for each individual and their complete donation and voting history.
3. Because the FEC data only includes an individual's first name, last name, and zip code, we cannot be certain that duplicates in the PA voter file are associated with a unique donation history. As a result, we drop all duplicate records from our data. Of the roughly 8 million voter files in the data, around 400,000 (or 4.6 percent) are duplicates. Again, duplicates were defined as multiple records with the same first name and last name living in the same zip code.
4. Any voter file that did not have any donation records were filled with zeros for donation amount over time. To handle NA values for the other numeric or integer variables, we replaced with the mean value for that variable. To handle NA values for categorical variables, we replaced with the mode value for that variable.
5. Random Forest can only handle factors with 53 or fewer levels. So, we reduced the number of categories for our factors to be fewer than 53.

2 Exploratory Data Analysis

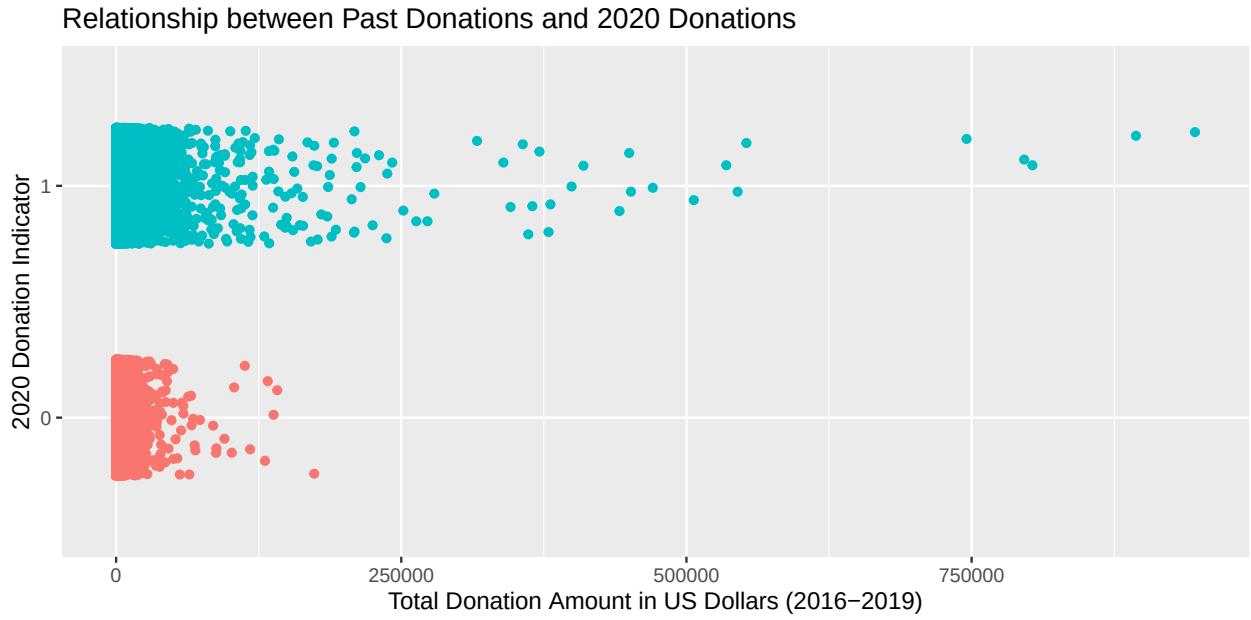
There are 8,169,268 unique voters included in the final cleaned data. For each voter, the data includes 70 features, which are described in greater detail in the Appendix. There are 121,851 PA voters who donated to a federal election in 2020 (i.e., our target variable), which is about 1.49 percent of voters.

The above correlation matrix shows that various features are associated with donating in 2020. Past donation behavior, county-level median income, county-level percent college educated, donor's age, and donor's predicted race as white are all positive associated with donating to a federal campaign in 2020. Donor's predicted race as black is negatively associated with donating to a federal campaign in 2020.

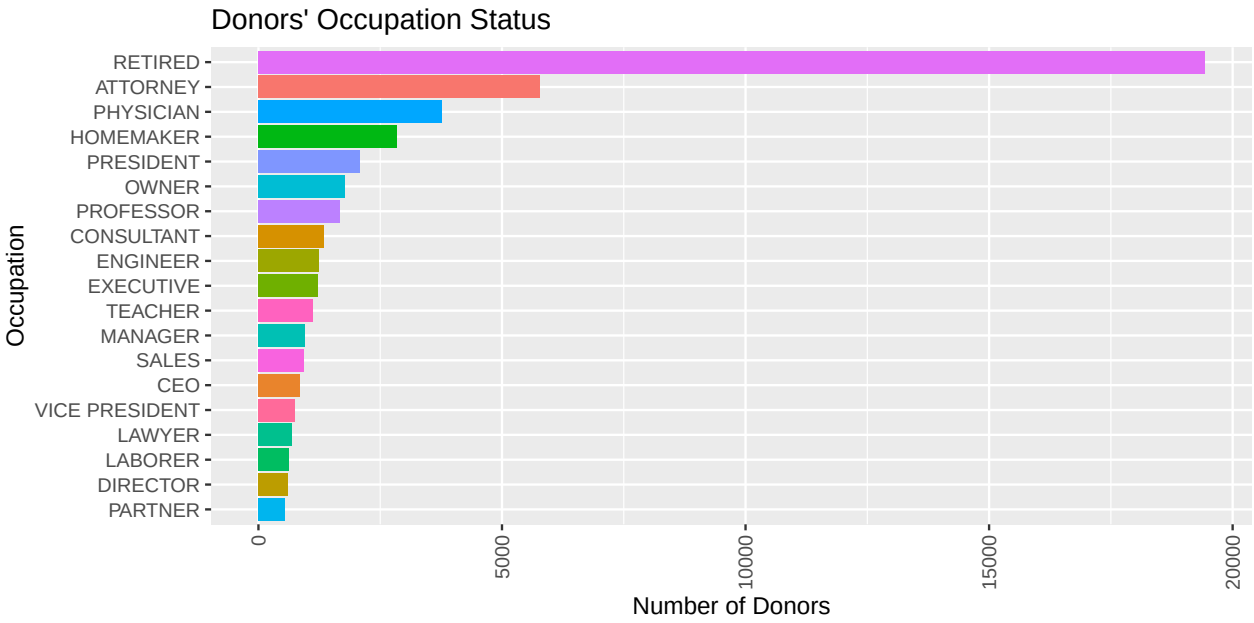
2020 Federal Donation Activity in Pennsylvania



The above plot shows the distribution of 2020 donors across PA. As expected, donors are heavily located in urban areas with higher populations like Philadelphia and Pittsburgh.



The above plot shows the relationship between an individual's total donation amount between the years 2016 and 2019 and whether an individual made a donation in 2020. The plot demonstrates that individuals who donated more money in the past are more likely to give in 2020.



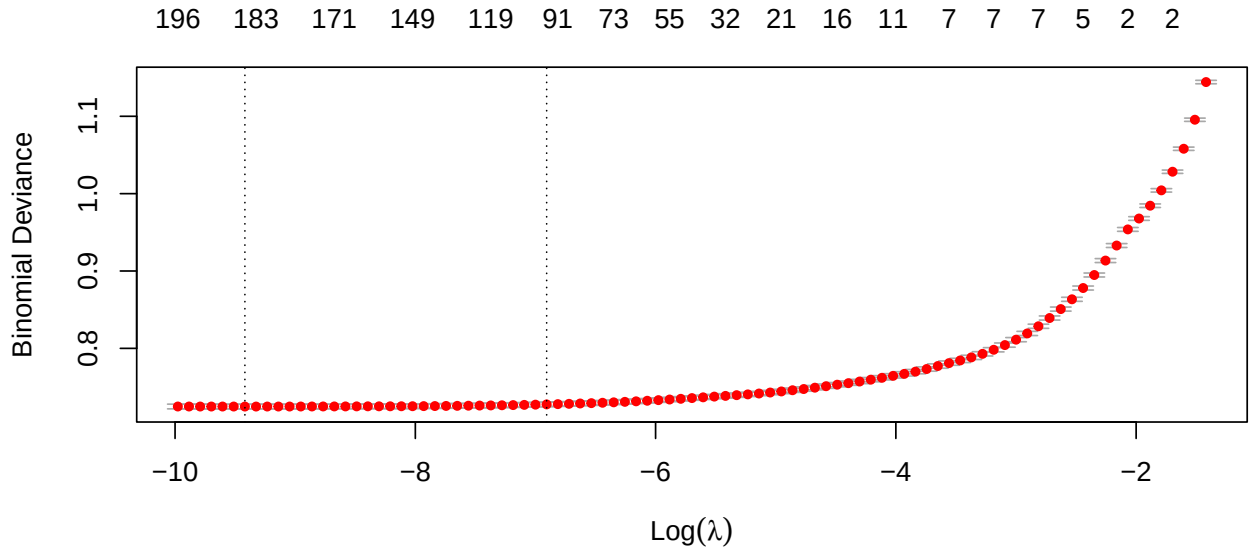
The above plot shows the listed occupation of donor's included in the data. Other than "none listed," the most popular occupation for donors is retired, which is then followed by attorneys and physicians.

3 Data Analysis

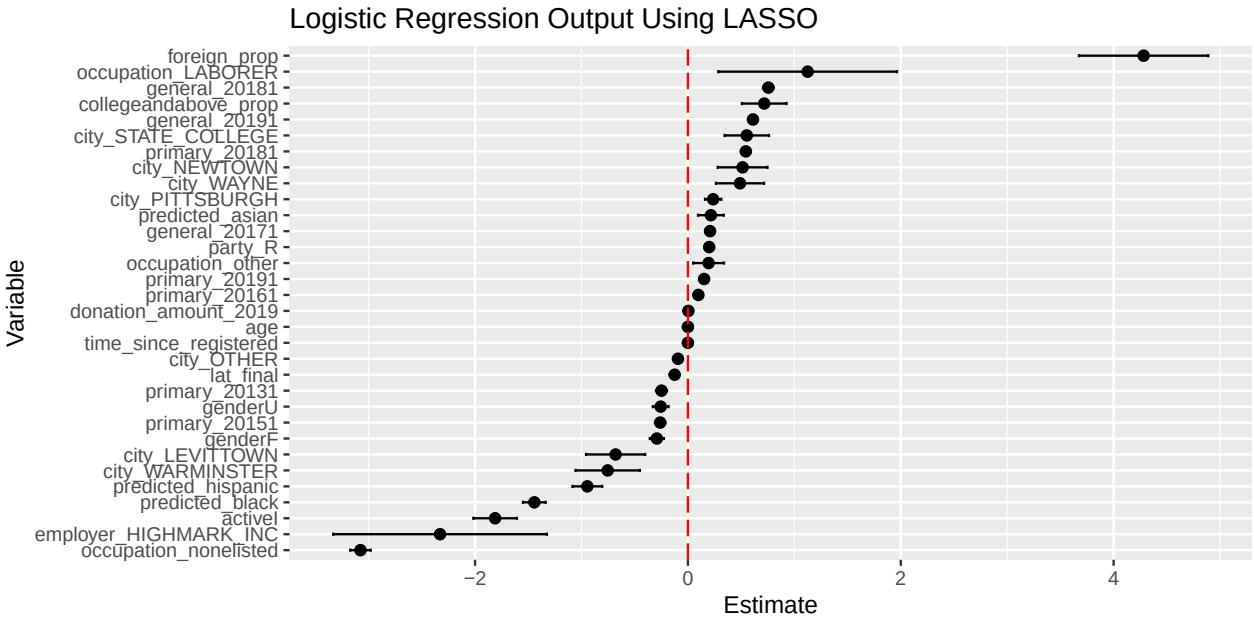
For computational efficiency, we take a weighted random sample of the full dataset. After splitting the data into training and testing sets, the remaining sample for each set is 135,000. For both training and testing data, there are 35,000 instances of 2020 donors and 100,000 instances of non-donors.

3.1 Logistic Regression Using LASSO

First, we implement logistic regression using LASSO. Logistic regression uses a set of predictors to capture the probability that an event occurs. For this analysis, we have 65 predictors (i.e., features) and the target variable is a binary indicator for whether an individual donated money to a federal campaign in 2020. To avoid overfitting, we implement LASSO (i.e., "Least Absolute Shrinkage and Selection Operator"), which uses k-fold cross-validation and a penalty term for additional features. The goal of LASSO is to return a subset of important variables in order to select a parsimonious model.



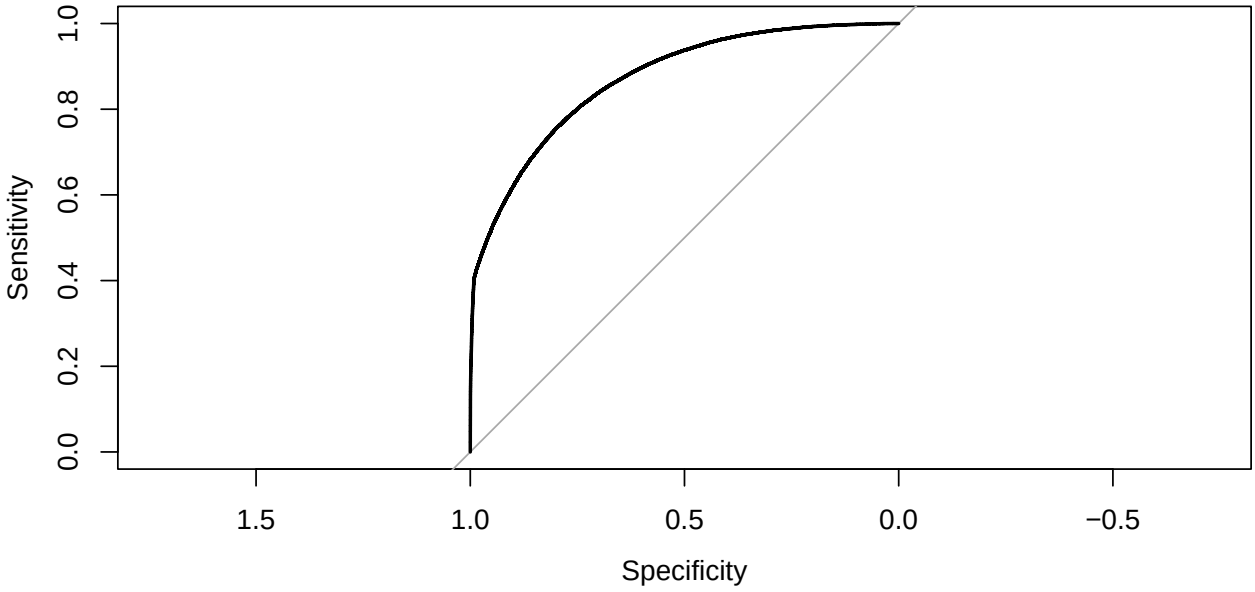
The above plot shows the performance of the LASSO logistic regression model as measured by deviance across various values for lambda (i.e., the penalization hyperparameter). The plot shows that adding additional predictive features gradually reduces the deviance. Around log(lambda) = -5.5, however, the model barely reduces its deviance despite adding significantly more variables. As a result, we manually set the lambda term to equal 0.00409 (i.e., exponential(-5.5)), which returns 36 variables.



The above plot shows the logistic regression coefficients and their confidence intervals using LASSO. As the plot shows, voters living in counties with higher foreign born populations and higher levels of college education are positively associated with donations in 2020. Voters who are laborers, who voted in 2018 and

2019, and who live in State College, Newtown, Wayne, and Pittsburgh are also more likely to make donations in 2020. Registered Republicans and Asian Americans are all more likely to donate.

On the other hand, individuals who voted in the 2013 and 2015 primaries, women, Hispanics, African Americans, inactive voters, voters employed by Highmark Inc. and individuals with no occupation listed are less likely to make donations. Voters living in Levittown and Warminster were also less likely to make donations.




```
##
## Call:
## roc.default(response = data_test_analysis_1_dummy$donation_2020_binary,     predictor = fit.lasso.pr
##
## Data: fit.lasso.pred in 100000 controls (data_test_analysis_1_dummy$donation_2020_binary 0) < 35000
## Area under the curve: 0.87
```

The above ROC curve informs how well the model is able to distinguish between classes. The curve shows the proportion of classified true positives (y-axis) and false positives (x-axis). The curve is generated by varying the probability threshold for classification. This curve can be used to select an optimal threshold. A classification threshold of 0.36 produces a False Positive rate of less than .1 and a True Positive rate as high as possible.

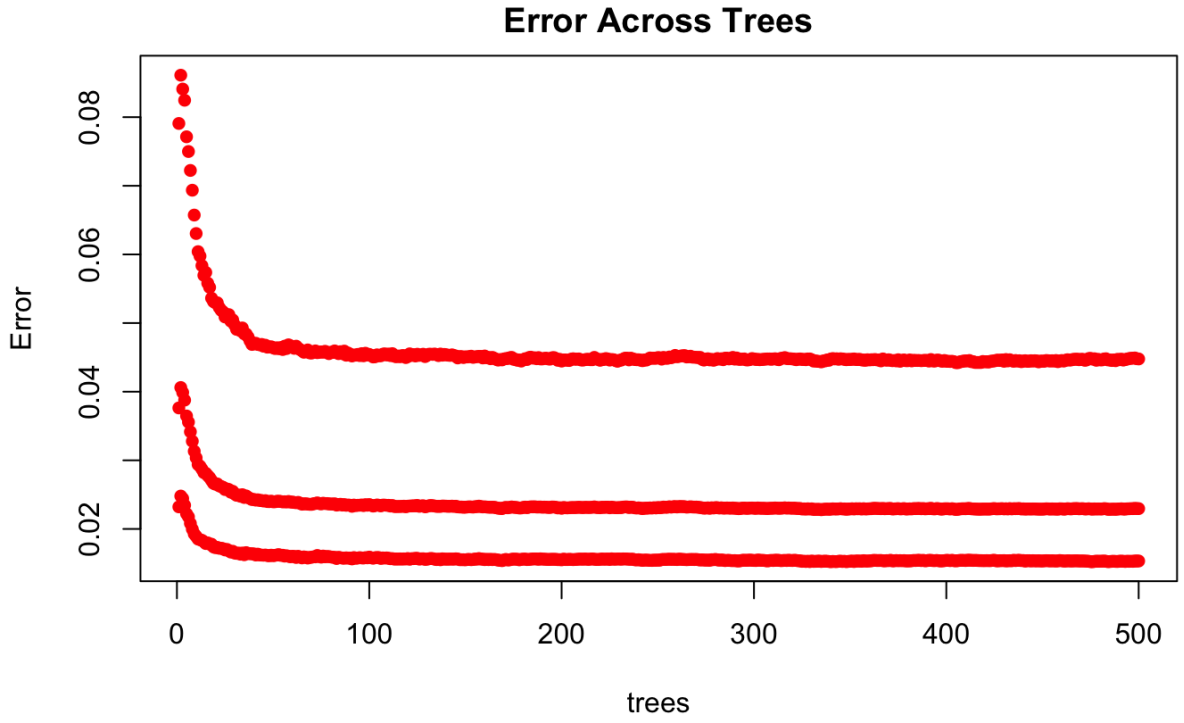
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 90030 13328
##          1  9970 21672
##
##                Accuracy : 0.827
##                  95% CI : (0.825, 0.829)
##     No Information Rate : 0.741
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.536
##
##  Mcnemar's Test P-Value : <2e-16
##
```

```
##             Sensitivity : 0.619
##             Specificity : 0.900
##          Pos Pred Value : 0.685
##          Neg Pred Value : 0.871
##              Prevalence : 0.259
##          Detection Rate : 0.161
##    Detection Prevalence : 0.234
##       Balanced Accuracy : 0.760
##
##        'Positive' Class : 1
##
```

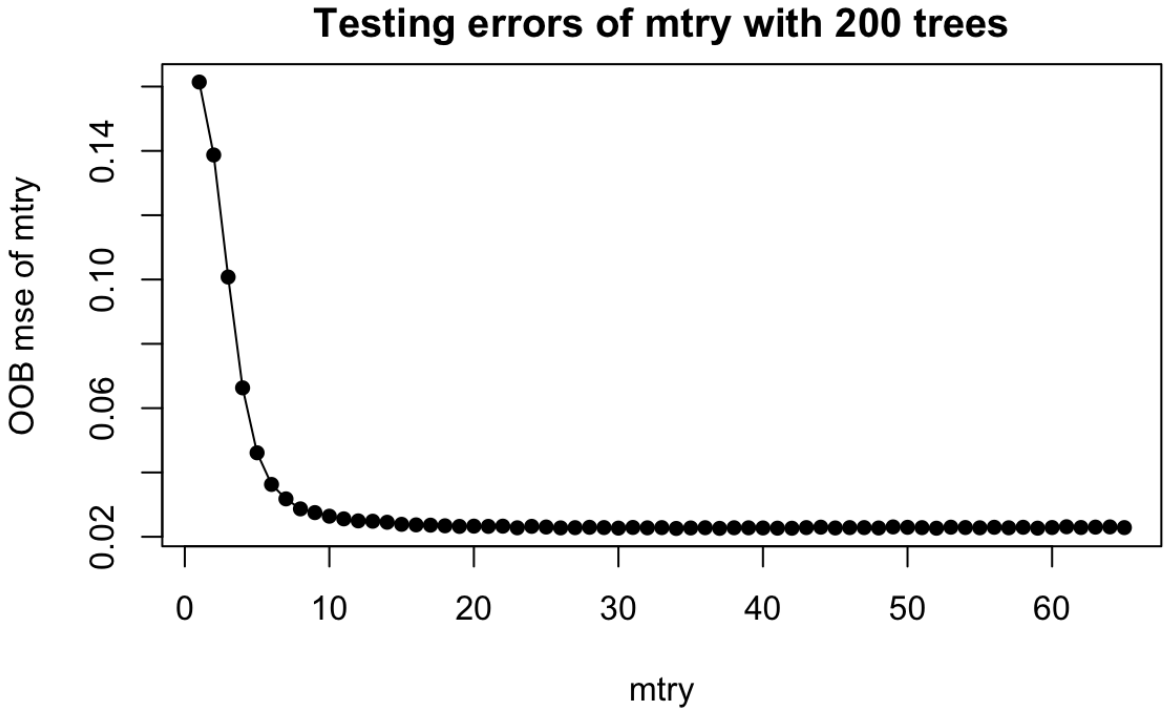
Using held out testing data and a threshold of 0.36, the confusion matrix shown above shows that the LASSO logistic regression model is able to correctly classify 82.7 percent of cases (i.e., 17.3 percent testing error). This is an improvement on the no information rate, which is 0.741. More importantly, the model correctly identifies positive classes (i.e., 2020 donors) at 62 percent. The model correctly identifies negative classes (i.e., non-2020 donors) at 90 percent.

3.2 Random Forest

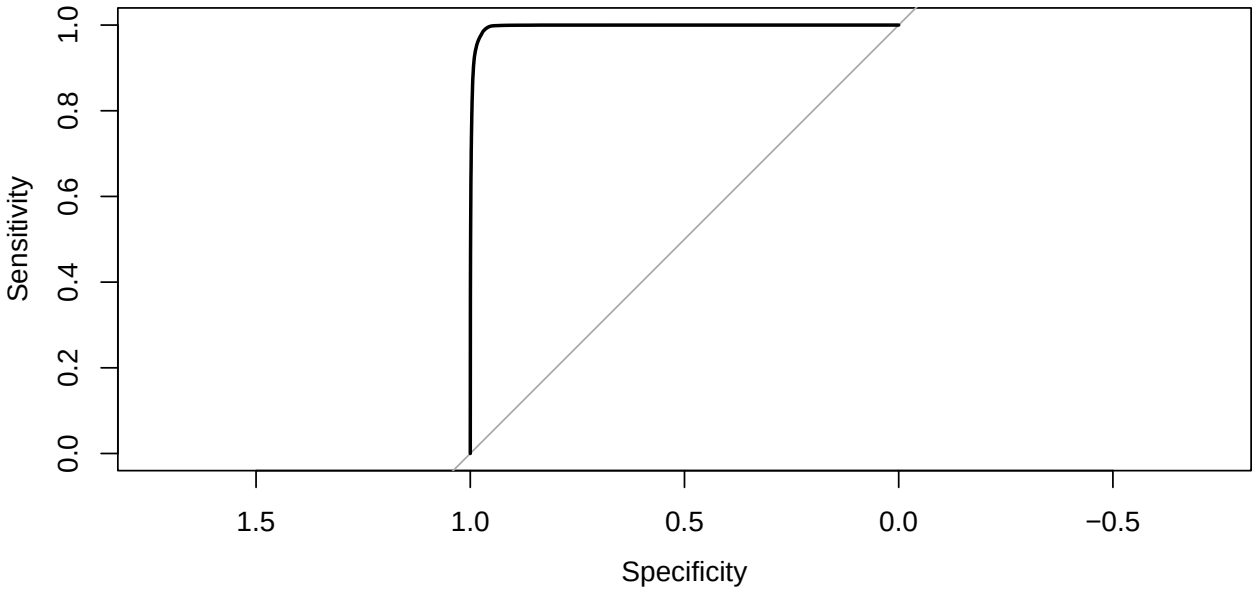
Second, we implement random forest, which is a machine learning algorithm that combines the logic of bagging and bootstraps. Bagging means combining multiple uncorrelated estimators and a bootstrap involves sampling with replacement from the underlying data to create slightly different samples across estimators. The core idea behind random forest is combining numerous decision trees trained on bootstrap samples for a random subset of features at each split for each tree. The model parameters are tuned using out of bag observations for each tree.



First, we tune the number of trees to include in the random forest model. The above plot shows out of bag misclassification rate for each class across numerous trees. The results show that the model's error rate stables off around 200 trees. As a result, we use 200 trees for our random forest algorithm.



Having already optimized the number of trees, we next tune the number of features to randomly sample at each split for each tree. The above plot shows the out of bag misclassification rate across different feature sizes to sample. The results shows that the model's error rate stables off around 20 features. Our final model thus sets mtry to 20 and ntree to 200 for its hyperparameters.




```
##
## Call:
## roc.default(response = data_test_analysis_1$donation_2020_binary,     predictor = predict.rf[, 2], pl
##
## Data: predict.rf[, 2] in 100000 controls (data_test_analysis_1$donation_2020_binary 0) < 35000 cases
```

```
## Area under the curve: 0.997
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 90429    16
##          1  9571 34984
##
##                Accuracy : 0.929
##                  95% CI : (0.928, 0.93)
##     No Information Rate : 0.741
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.83
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.000
##             Specificity : 0.904
##          Pos Pred Value : 0.785
##          Neg Pred Value : 1.000
##              Prevalence : 0.259
##          Detection Rate : 0.259
##    Detection Prevalence : 0.330
##       Balanced Accuracy : 0.952
##
##        'Positive' Class : 1
##
```

Finally, we assess model performance on held out test data. Looking to the above ROC curve, a classification threshold of 0.032 produces a False Positive rate of less than .1 and a True Positive rate as high as possible.

Using held out testing data and a threshold of 0.032, the confusion matrix shown above shows that random forest is able to correctly classify 92.9 percent of cases (i.e., 7.1 percent testing error). This is a significant improvement on the no information rate, which is 0.741. Incredibly, the model correctly identifies positive classes (i.e., 2020 donors) at 99.99 percent. Of the 35,000 instances of 2020 donors, random forest correctly classifies 34984 as donors. The model correctly identifies negative classes (i.e., non-2020 donors) at 90.4 percent.

4 Conclusion

As a whole, our analyses show that random forest and LASSO logistic regression are very effective at predicting campaign contribution behavior. Comparing logistic regression and random forest, however, reveals that random forest is the superior predictive algorithm. While LASSO logistic regression correctly classifies 82.7 percent of test cases, random forest correctly classifies 92.8 percent of test cases. More importantly, random forest is a remarkable 38 percentage points higher when it comes to positive classification (i.e., 61.9 percent versus 99.9 percent).

These types of analyses have the power to dramatically reshape the nature of political campaigning. If campaigns target their advertising toward voters who are more likely to donate, then candidates may be more likely to increase their fundraising capabilities. Future iterations of this project will look to expand the data beyond PA and will also aim to make future predictions as test cases, rather than held out data from the same year.

5 References

1. Broockman, David, and Joshua Kalla. 2016. "Durably Reducing Transphobia: A Field Experiment on Door-to-Door Canvassing." Science (American Association for the Advancement of Science) 352: 220-24.
2. Gerber, Alan S., and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." The American political science review 94: 653-63.

6 Appendix

The below tables list each of the included variables and information on their distribution. For numeric variables, we include a voter's location as latitude and longitude, their predicted race, the amount of money they donated annually from 2000 to 2019, their age, and the time since they initially registered to vote. We also include county-level information on total votes in a county, Trump 2016 vote share, total population, median household income, percent white, percent black, percent Hispanic, percent foreign born, and percent college educated.

For categorical factors, we include a voter's gender, whether the voter is an active voter or inactive, their registered political party, their home city, their voting history, their occupation, their employer, and a binary indicator for whether they donated in 2020.

	Data Type	Variable	Mean	Standard Deviation
1	numeric	lat_final	40.44	0.52
2	numeric	long_final	-76.99	1.93
3	numeric	predicted_white	0.79	0.32
4	numeric	predicted_black	0.10	0.24
5	numeric	predicted_hispanic	0.06	0.19
6	numeric	predicted_asian	0.03	0.13
7	numeric	predicted_other	0.02	0.05
8	numeric	donation_amount_2000	0.91	144.47
9	numeric	donation_amount_2001	0.17	22.69
10	numeric	donation_amount_2002	0.25	33.34
11	numeric	donation_amount_2003	0.93	66.62
12	numeric	donation_amount_2004	2.07	492.36
13	numeric	donation_amount_2005	1.04	92.26
14	numeric	donation_amount_2006	1.90	291.66
15	numeric	donation_amount_2007	1.35	104.48
16	numeric	donation_amount_2008	3.19	435.17
17	numeric	donation_amount_2009	1.49	353.06
18	numeric	donation_amount_2010	2.63	246.59
19	numeric	donation_amount_2011	1.63	157.50
20	numeric	donation_amount_2012	4.53	523.45
21	numeric	donation_amount_2013	1.36	132.86
22	numeric	donation_amount_2014	2.12	269.49
23	numeric	donation_amount_2015	2.84	201.13
24	numeric	donation_amount_2016	6.42	492.56
25	numeric	donation_amount_2017	3.43	270.96
26	numeric	donation_amount_2018	6.20	494.74
27	numeric	donation_amount_2019	4.67	261.25
28	numeric	totalvotes	290694.52	239638.46
29	numeric	trump_voteshare2016	0.49	0.17
30	numeric	total_population	593427.88	494490.33
31	numeric	white_prop	0.81	0.16
32	numeric	aa_prop	0.11	0.12
33	numeric	hispanic_prop	0.07	0.06
34	numeric	noncitizen_prop	0.03	0.02
35	numeric	foreign_prop	0.07	0.04
36	numeric	collegeandabove_prop	0.32	0.10
37	numeric	median_income	64043.25	14885.95
38	numeric	age (days)	18893.09	6862.34
39	numeric	time_since_registered	7113.92	6159.53
40	numeric	donation_2000_through_2007	8.63	830.54
41	numeric	donation_2008_through_2011	8.94	826.03
42	numeric	donation_2012_through_2015	10.85	867.46
43	numeric	donation_2016_through_2019	20.71	1187.79

Table 1: Summary Statistics (Numeric Variables)

	Data Type	Variable	Factor Levels
1	factor	gender	F: 3277955, M: 2785152, U: 1171390, emp: 934771
2	factor	active	A: 7619828, I: 549439
3	factor	political_party	D: 3772313, R: 3199458, NF: 831366, I: 104041
4	factor	city	Oth: 4642799, PHI: 968617, PIT: 488389, REA: 121211
5	factor	primary_2012	0: 7021681, 1: 1147587
6	factor	general_2012	1: 4136508, 0: 4032760
7	factor	primary_2013	0: 7258968, 1: 910300
8	factor	general_2013	0: 6798897, 1: 1370371
9	factor	primary_2014	0: 7129242, 1: 1040026
10	factor	general_2014	0: 5425306, 1: 2743962
11	factor	primary_2015	0: 6949836, 1: 1219432
12	factor	general_2015	0: 6355153, 1: 1814115
13	factor	primary_2016	0: 5417870, 1: 2751398
14	factor	general_2016	1: 5039462, 0: 3129806
15	factor	primary_2017	0: 6952414, 1: 1216854
16	factor	general_2017	0: 6277553, 1: 1891715
17	factor	primary_2018	0: 6780993, 1: 1388275
18	factor	general_2018	1: 4373075, 0: 3796193
19	factor	primary_2019	0: 6679810, 1: 1489458
20	factor	general_2019	0: 5694848, 1: 2474420
21	factor	occupation_mode	NON: 8011357, Oth: 72392, RET: 36782, ATT: 6803
22	factor	employer_mode	NON: 8016837, Oth: 94376, RET: 30859, SEL: 15422
23	factor	donation_2020_binary	0: 8047417, 1: 121851

Table 2: Summary Statistics (Categorical Variables)