Where's the Money? Campaign Finance and Election Outcomes in the 2016 U.S. Congressional Races

By Sarah Scott
MACS 33002 Project I
February 12, 2023

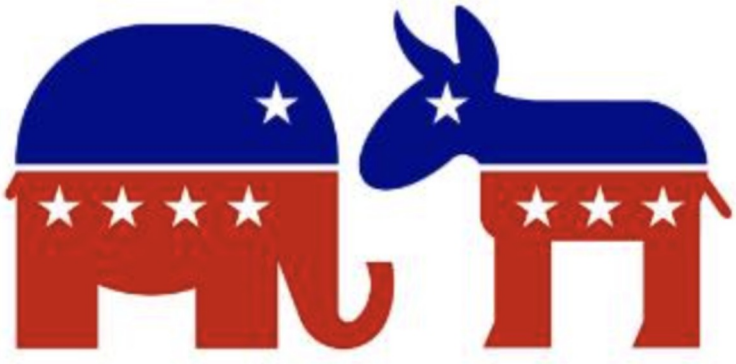
Project Overview

Interest

- Voters, candidates, and news stations often ask what factors go into a winning campaign

Goal & Project Type

- Predict whether a candidate is a “Winner” or “Loser” based on campaign finance data
- Binary Classification



Data Processing

Feature Review and Engineering

- Raw Data: 1814 candidates, 51 features
- Processed data: 1656 candidates, 20 features
  - 17 numeric
  - 3 categorical variables

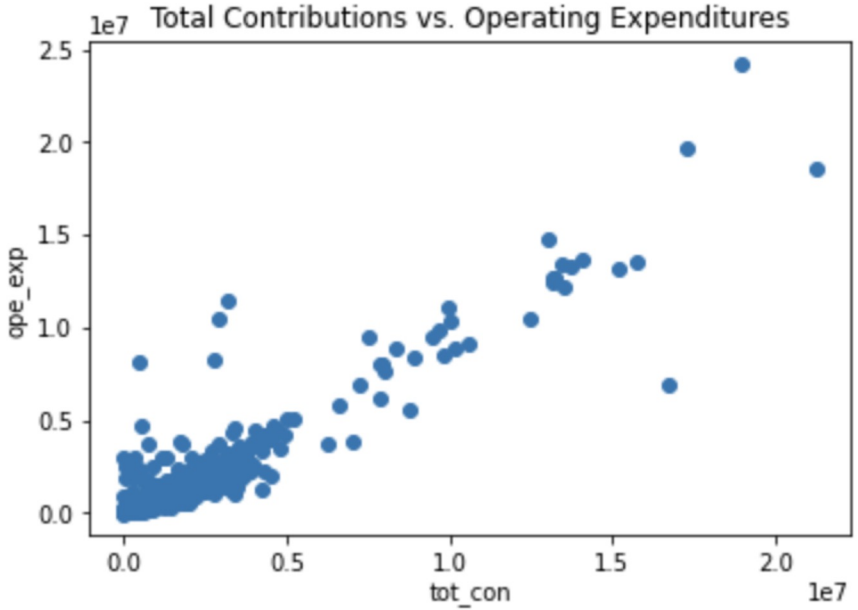
	can_id	can_nam	can_off	can_off_sta	can_off_dis	can_par_aff	can_inc_cha_ope_sea	can_str1	can_str2	can_cit	...	cas_on_han
0	H2GA12121	ALLEN, RICHARD W	H	GA	12.0	REP	INCUMBENT	2237 PICKENS RD	NaN	AUGUSTA	...	
1	H6PA02171	EVANS, DWIGHT	H	PA	2.0	DEM	CHALLENGER	PO BOX 6578	NaN	PHILADELPHIA	...	
2	H6FL04105	RUTHERFORD, JOHN	H	FL	4.0	REP	OPEN	3817 VICKERS LAKE DRIVE	NaN	JACKSONVILLE	...	
3	H4MT01041	ZINKE, RYAN K	H	MT	0.0	REP	INCUMBENT	409 2ND ST W	NaN	WHITEFISH	...	
4	H8CA09060	LEE, BARBARA	H	CA	13.0	DEM	INCUMBENT	409 13TH ST, 17TH FL	NaN	OAKLAND	...	

Data Processing (continued)

Data Summary

- Feature highlights
  - Contributions
  - Loans
  - Operating Expenditures
  - Length of Campaign
  - Missing Values

- Final Sample distribution: 1186 losers, 470 winners



Model 1: Decision Trees

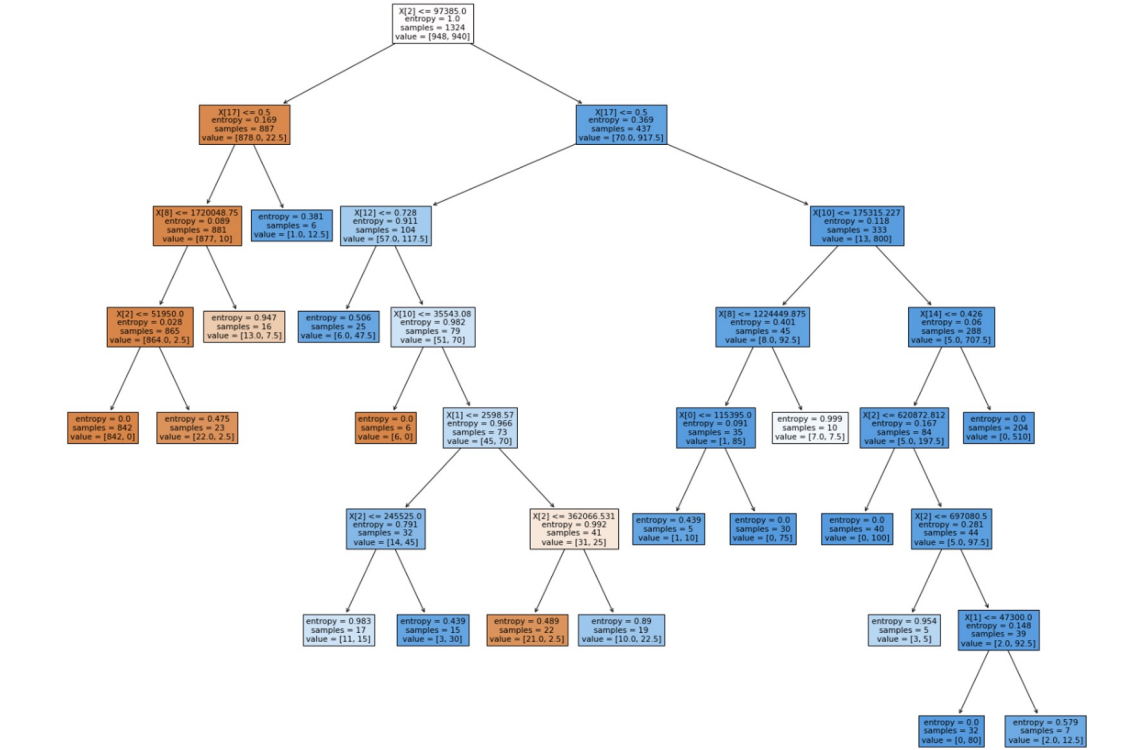
Decision Tree Classifier

- One-Hot Encoder for categorical variables
  - Challenger, Incumbent, or Open Race Status
  - House or Senate Race
  - Democrat, Republican, Other
- Training on 80% of data, Testing on 20% of data
- Weights: Winners at 2.5, Losers at 1
- Parameter Setting

Model 1: Results Analysis

Decision Tree Classifier

- Accuracy: 94.3%
- Top Features:
  - Other Committee Contributions (83%)
  - Incumbent Status (10%)
- Results Analysis
  - 19 prediction errors out of 332 candidates
  - Overpredicted that candidates won races
  - Over-reliant on PACs and Super PACS contribution data



Model 2: Random Forest Classifier

Random Forest Classifier

- Similar setup as decision tree classifier
- Parameter Settings
- Accuracy: 95.78%
- 14 prediction errors out of 332 candidates

Model 2: Results Analysis

	feature	importances
2	oth_com_con	0.184804
17	(Incumbent,)	0.126577
4	tot_con	0.112342
14	can_con_tot_con_ratio	0.103925
13	oth_con_tot_con_ratio	0.092812
8	tot_rec	0.079900
10	cas_on_han_clo_of_per	0.058828
9	ope_exp	0.055860
11	campaign_days_total	0.044882
0	ind_con	0.039824
12	ind_con_tot_con_ratio	0.028748
16	(Challenger,)	0.024326

Random Forest Classifier

- Top Features
  - Super PACs
  - Incumbent Status
  - Total Contributions
  - Contribution Ratios

- By generating multiple decision trees and then taking the prediction that most of the trees return, the resulting model weights more features evenly rather than relying heavily on one feature.

- Similar to the other model, we see that the model mislabeled losers as winners.

Conclusion

- Important features: PACs and Super PACs contributions, as well as incumbent status

- Limitations:
  - Effects of geography on outcome
  - Challenging to use election data because of unbalanced samples