Quantifying the Return on Investment (ROI) of Campaign Contributions

Hope Mullins

University of Central Florida

July 7, 2025

Original Database

- Database on Ideology, Money in Politics, and Elections (DIME) (Version 4.0)
- 1979 – 2024
- Three Tables
  - Candidate/Recipient Table: 64 Columns
  - Contributions Table: 45 Columns
  - Donor Table: 43 Columns

Filtering

- DuckDB
- candDB
  - `seat` = federal:house, `cycle` = 2000 – 2024, `bonica_rid` not null
- contribDB
  - shared `bonica_rids` with candDB, `seat` = federal:house, `date` = 1998 – 2024
- donorDB
  - shared `bonica_cids` with contribDB

Variable Cleaning

- Candidate/Recipient Table
- Originally: 64 Variables
  - Voting Records
  - Fundraising Statistics
  - Election Outcomes
  - Candidate Characteristics
- After Filtering: 19 Variables

Variable Cleaning (Continued)

- Candidate/Recipient Table
  - Used: `cycle`, `bonica_rid`, `party`, `ico_status`, `num_givers`, `ind_exp_support`, `ind_exp_oppose`, `gen_vote_pct`, `gwinner`, `district_pres_vs`
  - Not Used: `bonica_cid`, `district`, `total_receipts`, `total_disbursements`, `total_indiv_contribs`, `total_unitemized`, `total_pac_contribs`, `total_party_contribs`, `total_contribs_from_candidate`

Variable Cleaning (Continued)

- Contributions Table
- Originally: 45 Variables
  - Contributor
  - Geographic
  - Financial
- After Filtering: 9 Variables

Variable Cleaning (Continued)

- Contributions Table
  - Used: `cycle`, `transaction_type`, `date`, `amount`, `bonica_rid`, `contributor_type`, `is_corp`
  - Not Used: `bonica_cid`, `election_type`

Variable Cleaning (Continued)

- Donor Table
- Originally: 43 Variables
  - Contributor
  - Geographic
  - Redundant Columns
- Dropped Table

Null Cleaning

	value
total_rows	97397007.0
cycle_nulls	0.0
transaction_type_nulls	14366.0
amount_nulls	0.0
date_nulls	0.0
bonica_cid_nulls	0.0
contributor_type_nulls	378.0
is_corp_nulls	94885587.0
bonica_rid_nulls	0.0
election_type_nulls	65226414.0
party_nulls	0.0
district_nulls	0.0
ico_status_nulls	0.0
num_givers_nulls	0.0
total_receipts_nulls	0.0
total_disbursements_nulls	0.0
total_indiv_contribs_nulls	0.0
total_unitemized_nulls	0.0
total_pac_contribs_nulls	0.0
total_party_contribs_nulls	0.0
total_contribs_from_candidate_nulls	0.0
ind_exp_support_nulls	0.0
ind_exp_oppose_nulls	0.0
gen_vote_pct_nulls	42154652.0
gwinner_nulls	40909810.0
district_pres_vs_nulls	107895.0

Null Cleaning (Continued)

- Joined on `cycle` and `bonica_rid`
- $97,397,007$ rows
- Kept only general elections: $13,933,493$
- Dropped NULL `gwinner` and `gen_vote_pct`: $9,447,843$
- Dropped all NULLs besides `is_corp`: $9,430,440$
- Converted `is_corp` NULLs to 0 and value 'corp' to 1

Converting

☆ CPI-based deflators

Converting

- ☆ CPI-based deflators
- ☆ Days before election

Converting

- ☆ CPI-based deflators
- ☆ Days before election
- ☆ Binary/one-hot encoding: `contributor_type`, `ico_status`, `gwinner`

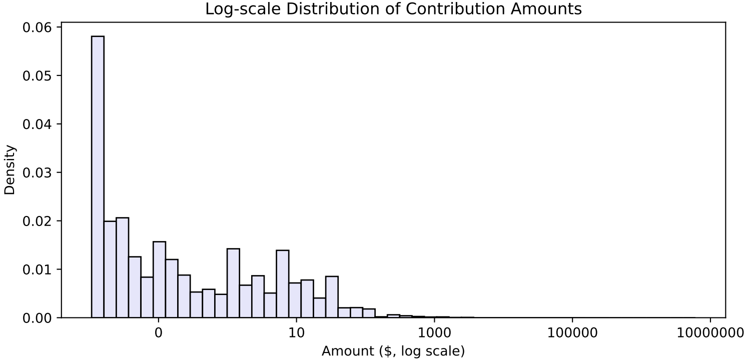
Converting

- ☆ CPI-based deflators
- ☆ Days before election
- ☆ Binary/one-hot encoding: `contributor_type`, `ico_status`, `gwinner`
- ☆ Frequency-encoded: `transaction_type`

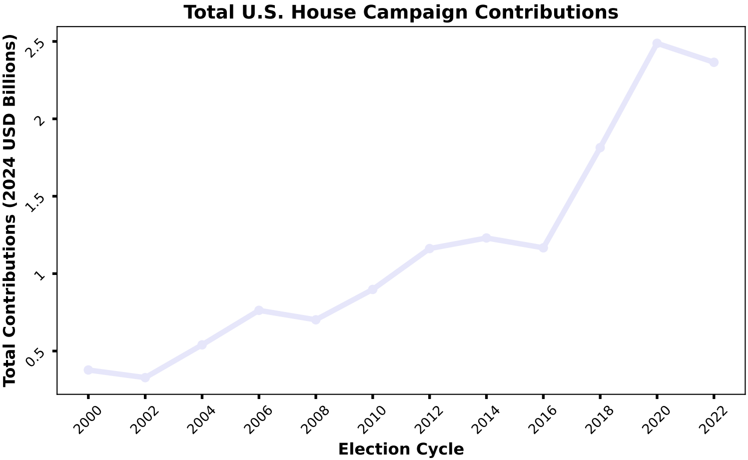
Converting

- ☆ CPI-based deflators
- ☆ Days before election
- ☆ Binary/one-hot encoding: `contributor_type`, `ico_status`, `gwinner`
- ☆ Frequency-encoded: `transaction_type`
- ☆ Ready for aggregation and modeling

Contribution Amount Distribution



Total Contribution Spending Trend



Aggregating

- Constant returns across time?
- Automate by cutoff dates
- New/adjusted variables: `n_contribs_Xd`, `avg_tx_freq_Xd`, `indiv_mill_Xd`, `comm_mill_Xd`, `corp_mill_Xd`, `party_D`, `party_R`, `party_Other`
- Aggregate up to a particular date
- Save to files

Aggregation Loop

```
# Aggregate and save as Parquet
for X in cutoffs:
    print(f"Aggregating for {X} days before election...")

    con.execute(f"""
        CREATE OR REPLACE TABLE agg_rid_{X}d AS
        SELECT
          bonica_rid,
          cycle,
          COUNT(*)                                         AS n_contribs_{X}d,
          AVG(tx_type_freq)                                AS avg_tx_freq_{X}d,
          SUM((1 - contrib_type) * amount) / 1e6           AS indiv_mill_{X}d,
          SUM(contrib_type * (1 - is_corp) * amount) / 1e6 AS comm_mill_{X}d,
          SUM(contrib_type * is_corp * amount) / 1e6       AS corp_mill_{X}d,
          MAX(num_givers)                                  AS num_givers,
          MAX(total_receipts)                              AS total_receipts,
          MAX(total_disbursements)                         AS total_disbursements,
          MAX(total_indiv_contribs)                        AS total_indiv_contribs,
          MAX(total_unitemized)                            AS total_unitemized,
          MAX(total_pac_contribs)                          AS total_pac_contribs,
          MAX(total_party_contribs)                        AS total_party_contribs,
          MAX(total_contribs_from_candidate)               AS total_self_contribs,
          MAX(ind_exp_support)                             AS ind_exp_support,
          MAX(ind_exp_oppose)                              AS ind_exp_oppose,
          MAX(gen_vote_pct)                                AS gen_vote_pct,
          MAX(district_pres_vs)                            AS pres_margin,
          MAX(CASE WHEN party = 100  THEN 1 ELSE 0 END)       AS party_D,
          MAX(CASE WHEN party = 200 THEN 1 ELSE 0 END)       AS party_R,
          MAX(CASE WHEN party NOT IN (100, 200) THEN 1 ELSE 0 END)
                                                           AS party_Other,
          MAX(is_incumbent)                                AS incumbent,
          MAX(gwinner)                                     AS won_general
        FROM house
        WHERE days_before >= {X}
        GROUP BY bonica_rid, cycle;
    """)
```

Logistic Model

- Features: n_contribs_Xd, avg_tx_freq_Xd, indiv_mill_Xd, comm_mill_Xd, corp_mill_Xd, num_givers, ind_exp_support, ind_exp_oppose, pres_margin, party_R, party_Other, incumbent
- Log-transformed money variables
- Training 75%; testing 25%
- Standardized data
- MLE logistic regression

Logistic Model Metrics

Days	AUC	Accuracy	Recall (1)	Recall (0)
360	0.899	0.894	0.917	0.801
240	0.891	0.865	0.868	0.858
120	0.915	0.885	0.858	0.915
60	0.932	0.896	0.866	0.934
30	0.934	0.888	0.845	0.937
14	0.933	0.894	0.858	0.936
7	0.931	0.887	0.857	0.920
1	0.942	0.899	0.859	0.943

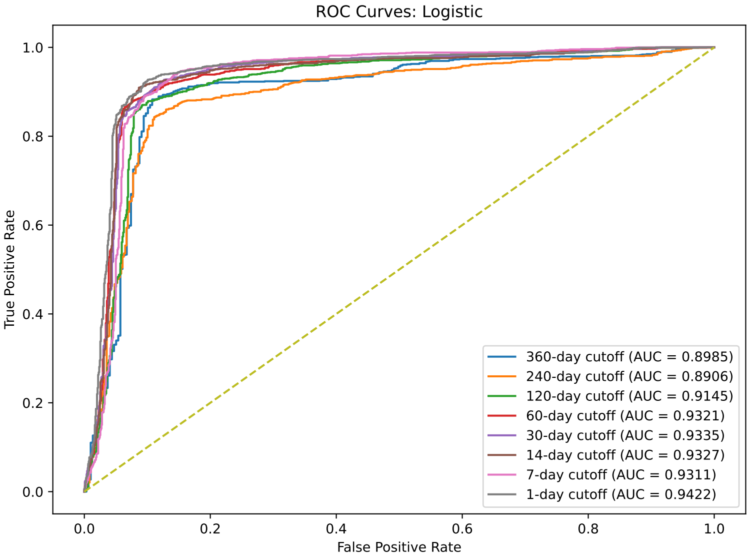
Logistic Model ROI

Days	Total ROI β	Odds-Ratio
360	–0.102	0.90
240	-0.035	0.97
120	0.341	1.41
60	0.672	1.96
30	0.737	2.09
14	0.831	2.30
7	0.699	2.01
1	0.673	1.96

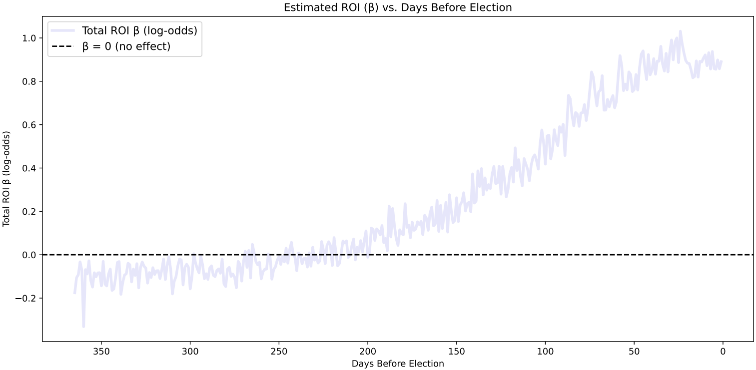
Logistic Model By-Type ROIs

Days	Ind β	Ind OR	Comm β	Comm OR	Corp β	Corp OR
360	0.070	1.07	–0.040	0.96	–0.132	0.88
240	0.179	1.20	–0.106	0.90	–0.108	0.90
120	0.263	1.30	0.031	1.03	0.047	1.05
60	0.422	1.52	–0.117	0.89	0.367	1.44
30	0.439	1.55	–0.199	0.82	0.498	1.65
14	0.482	1.62	–0.300	0.74	0.649	1.91
7	0.563	1.76	–0.341	0.71	0.478	1.61
1	0.526	1.69	–0.373	0.69	0.521	1.68

Logistic ROCs

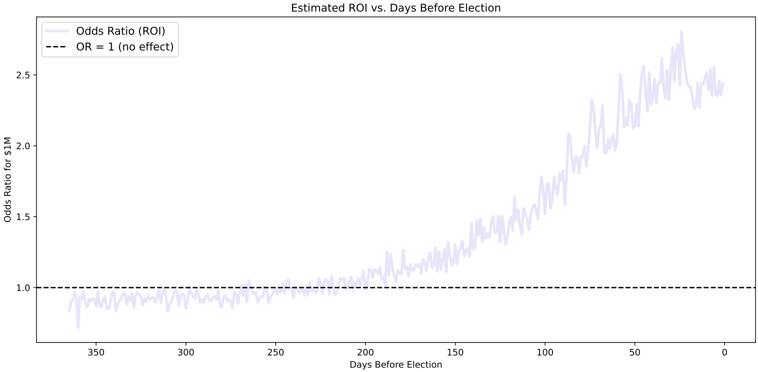


ROI Trend



Estimated ROI (β) vs. Days Before Election
Total ROI β (log-odds)
β = 0 (no effect)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
350
300
250
200
150
100
50
0
Days Before Election
Total ROI β (log-odds)

Odds-Ratio Trend



Win-Prob GAM Model

- ➔ Same features as before
- ➔ Log-transformed to reduce skew
- ➔ Standardized data
- ➔ Training 75%; testing 25%
- ➔ GAM specifications

Win-Prob GAM Metrics

Days	AUC	Accuracy	Recall (1)	Recall (0)
360	0.914	0.896	0.939	0.724
240	0.911	0.868	0.894	0.799
120	0.938	0.887	0.891	0.882
60	0.955	0.904	0.912	0.895
30	0.960	0.904	0.909	0.898
14	0.966	0.911	0.916	0.905
7	0.963	0.909	0.928	0.887
1	0.965	0.912	0.918	0.905

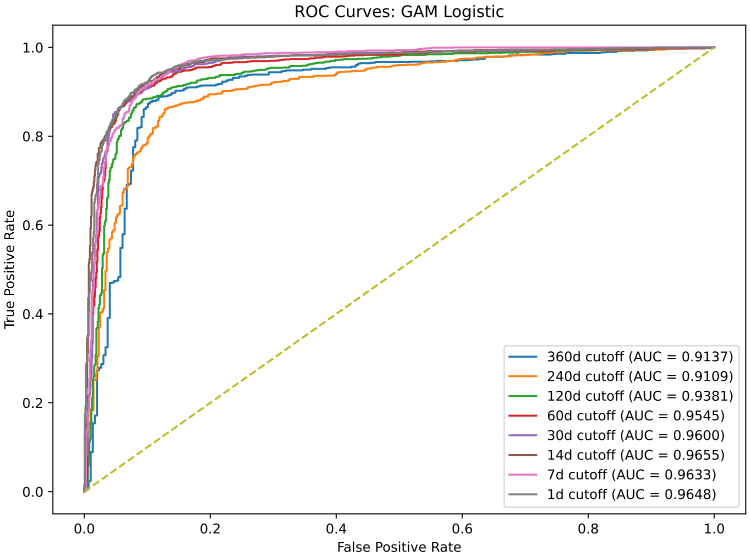
Win-Prob GAM ROI

Days	Total ROI β	Odds-Ratio
360	–0.285	0.75
240	0.072	1.07
120	0.198	1.22
60	-0.107	0.90
30	0.371	1.45
14	0.033	1.03
7	0.308	1.36
1	0.503	1.65

Win-Prob GAM By-Type ROIs

Days	Ind β	Ind OR	Comm β	Comm OR	Corp β	Corp OR
360	-0.037	0.96	–0.399	0.67	–0.529	0.59
240	0.471	1.60	–0.111	0.90	–0.275	0.76
120	0.221	1.25	0.228	1.26	0.140	1.15
60	-0.737	0.48	0.699	2.01	-0.119	0.89
30	0.039	1.04	0.320	1.38	0.868	2.38
14	-0.065	0.94	–0.252	0.78	0.580	1.79
7	0.134	1.14	–0.129	0.88	1.210	3.35
1	0.318	1.37	–0.167	0.85	1.795	6.02

Win-Prob GAM ROCs



Linear Model

- Same features as before
- Log-transformed
- Training 75%; testing 25%
- Ordinary Least Squares Pipeline

Linear Model (Continued)

- Metrics/coefficients
- Standardized feature importance and unstandardized ROI
- Marginal effect
- Regularization (Lasso or Ridge)

Linear Metrics

Days	R^2	CV R^2	RMSE	MAE
360	0.410	0.393	13.672	9.715
240	0.468	0.461	14.343	10.185
120	0.555	0.560	14.372	10.208
60	0.580	0.585	14.232	10.135
30	0.587	0.596	14.168	9.980
14	0.603	0.599	13.899	9.837
7	0.609	0.601	13.926	9.987
1	0.595	0.601	14.163	10.157

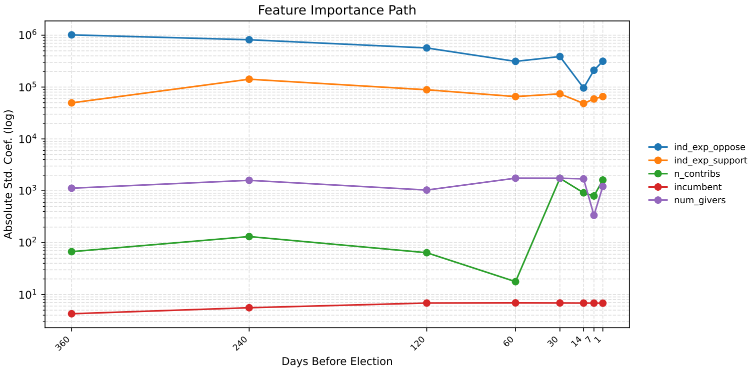
Regularized R-Squared

Days	Best Model (α)	R^2
360	R (100.000)	0.410
240	R (100.000)	0.468
120	L (0.1000)	0.556
60	R (17.7828)	0.580
30	L (0.0316)	0.588
14	L (0.0562)	0.604
7	R (31.6228)	0.609
1	L (0.0562)	0.595

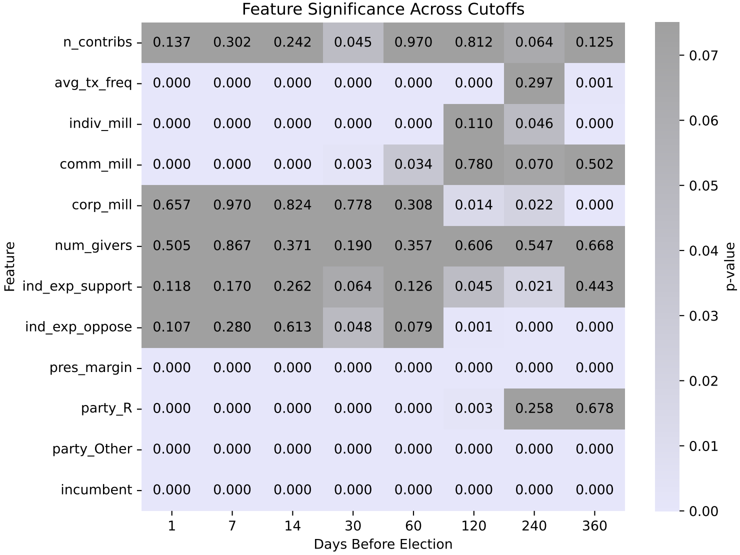
Linear ROI

Days	Total ROI β (pp per \$1M)	Δ (pp per \$1M)
360	–39.8554	–8.7519
240	–22.7274	–4.9645
120	-3.1081	-0.4435
60	1.9142	0.5872
30	5.7872	0.8855
14	6.0042	0.6739
7	5.3512	0.5154
1	6.3002	0.6224

Evolution Of Top Standardized Coefficients



Feature Statistical Significance



VS GAM Model

- Similar process to linear
- GAM specifications
- CV and hold-out metrics

VS GAM Metrics

Days	CV R^2	CV RMSE	CV MAE	Δ (pp per \$1M)
360	0.488	12.97	8.52	-1.60
240	0.538	13.34	8.97	-0.46
120	0.502	14.84	9.38	1.74
60	0.599	13.74	9.21	2.17
30	0.659	12.86	8.82	0.93
14	0.680	12.51	8.69	-0.19
7	0.681	12.53	8.72	-0.28
1	0.692	12.32	8.73	-0.23

VS GAM Metrics (Continued)

Days	HO R^2	HO RMSE	HO MAE
360	0.515	12.40	8.22
240	0.532	13.45	9.05
120	0.140	19.98	9.70
60	0.522	15.18	9.45
30	0.655	12.95	8.89
14	0.695	12.18	8.58
7	0.690	12.40	8.63
1	0.693	12.32	8.73

Win-Prob Metric Changes (GAM – MLE)

Days	Δ AUC	Δ Accuracy	Δ R (1)	Δ R (0)
360	0.015	0.002	0.022	-0.077
240	0.020	0.003	0.026	-0.059
120	0.023	0.002	0.033	-0.033
60	0.023	0.008	0.046	-0.039
30	0.026	0.016	0.064	-0.039
14	0.033	0.017	0.058	-0.031
7	0.032	0.022	0.071	-0.033
1	0.023	0.013	0.059	-0.038

VS Metric Changes (GAM – OLS)

Days	ΔR^2	$\Delta\mathrm{RMSE}$	$\Delta\mathrm{MAE}$
360	0.078	-0.702	-1.195
240	0.070	-1.003	-1.215
120	-0.053	0.468	-0.828
60	0.019	-0.492	-0.925
30	0.072	-1.308	-1.160
14	0.077	-1.389	-1.147
7	0.072	-1.396	-1.267
1	0.097	-1.843	-1.427

Summary

✔ Late money vs. early money

Summary

- ✔ Late money vs. early money
- ✔ Different contributor types behavior

Summary

- ✔ Late money vs. early money
- ✔ Different contributor types behavior
- ✔ GAM variants reveal nonlinearities

Summary

- ✔ Late money vs. early money
- ✔ Different contributor types behavior
- ✔ GAM variants reveal nonlinearities
- ✔ Variable importance/feature selection

Limitations

- ✘ Associational effects

Limitations

- ✘ Associational effects
- ✘ Convergence

Limitations

- ✘ Associational effects
- ✘ Convergence
- ✘ Class imbalance

Limitations

- ✘ Associational effects
- ✘ Convergence
- ✘ Class imbalance
- ✘ Arbitrary cutoffs

Limitations

- ✘ Associational effects
- ✘ Convergence
- ✘ Class imbalance
- ✘ Arbitrary cutoffs
- ✘ Additional tuning

Limitations

- ✘ Associational effects
- ✘ Convergence
- ✘ Class imbalance
- ✘ Arbitrary cutoffs
- ✘ Additional tuning
- ✘ Ease of automation

Conclusion

How do the magnitude and composition of campaign contributions affect a U.S. House candidate's probability of winning and expected vote share?

How can we use this information?

References I

- Bonica, Adam. *Database on Ideology, Money in Politics, and Elections: Public version 4.0 [Computer file]*. http://data.stanford.edu/dime. Stanford University Libraries, Stanford, CA, 2024.