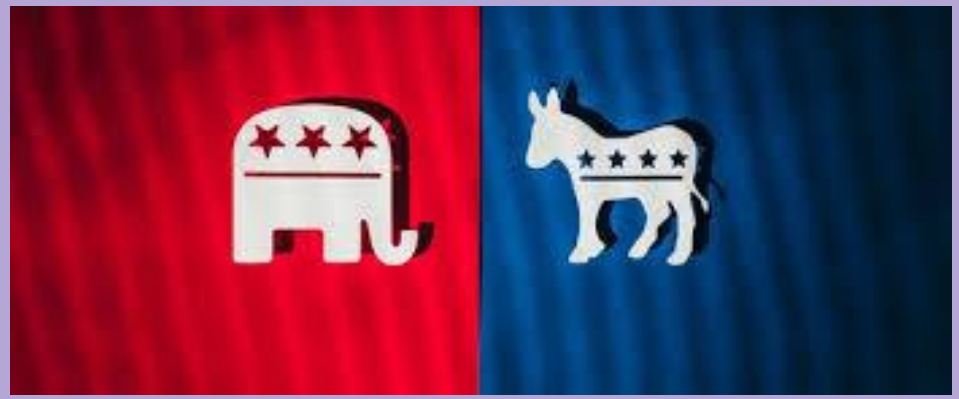
Leveraging Bellwether Counties for National Election Forecasting with Random Forest

Evan Shields

The Learning Problem

This project defines a supervised classification problem, where the model's task is to predict the binary outcome (Democratic or Republican win) of presidential elections in specific bellwether counties. By training on historical data and learning from patterns in voting behavior alongside demographic data, the model aims to generalize well to future election predictions.



Data Used

The model used county-level data, including historical voting results from past elections. Additionally, recent demographic data (age distribution, race, education, etc.) was added into the input data.

Sources:

- County Presidential Election Returns 2000-2020
  https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VOQCHQ
- County-level US demographic data, 1990-2020
  https://www.kaggle.com/datasets/glozab/county-level-us-demographic-data-1990-2020

countypres_2000-2020

year	state	state_po	county_name	county_fips	office	candidate	party	candidatevotes	totalvotes	version	mode
2000	ALABAMA	AL	AUTAUGA	1001	US PRESIDENT	AL GORE	DEMOCRAT	4942	17208	20220315	TOTAL
2000	ALABAMA	AL	AUTAUGA	1001	US PRESIDENT	GEORGE W. BUSH	REPUBLICAN	11993	17208	20220315	TOTAL
2000	ALABAMA	AL	AUTAUGA	1001	US PRESIDENT	RALPH NADER	GREEN	160	17208	20220315	TOTAL
2000	ALABAMA	AL	AUTAUGA	1001	US PRESIDENT	OTHER	OTHER	113	17208	20220315	TOTAL
2000	ALABAMA	AL	BALDWIN	1003	US PRESIDENT	AL GORE	DEMOCRAT	13997	56480	20220315	TOTAL
2000	ALABAMA	AL	BALDWIN	1003	US PRESIDENT	GEORGE W. BUSH	REPUBLICAN	40872	56480	20220315	TOTAL
2000	ALABAMA	AL	BALDWIN	1003	US PRESIDENT	RALPH NADER	GREEN	1033	56480	20220315	TOTAL
2000	ALABAMA	AL	BALDWIN	1003	US PRESIDENT	OTHER	OTHER	578	56480	20220315	TOTAL
2000	ALABAMA	AL	BARBOUR	1005	US PRESIDENT	AL GORE	DEMOCRAT	5188	10395	20220315	TOTAL
2000	ALABAMA	AL	BARBOUR	1005	US PRESIDENT	GEORGE W. BUSH	REPUBLICAN	5096	10395	20220315	TOTAL
2000	ALABAMA	AL	BARBOUR	1005	US PRESIDENT	RALPH NADER	GREEN	46	10395	20220315	TOTAL
2000	ALABAMA	AL	BARBOUR	1005	US PRESIDENT	OTHER	OTHER	65	10395	20220315	TOTAL
2000	ALABAMA	AL	BIBB	1007	US PRESIDENT	AL GORE	DEMOCRAT	2710	7101	20220315	TOTAL
2000	ALABAMA	AL	BIBB	1007	US PRESIDENT	GEORGE W. BUSH	REPUBLICAN	4273	7101	20220315	TOTAL
2000	ALABAMA	AL	BIBB	1007	US PRESIDENT	RALPH NADER	GREEN	52	7101	20220315	TOTAL
2000	ALABAMA	AL	BIBB	1007	US PRESIDENT	OTHER	OTHER	66	7101	20220315	TOTAL
2000	ALABAMA	AL	BLOUNT	1009	US PRESIDENT	AL GORE	DEMOCRAT	4977	17973	20220315	TOTAL
2000	ALABAMA	AL	BLOUNT	1009	US PRESIDENT	GEORGE W. BUSH	REPUBLICAN	12667	17973	20220315	TOTAL
2000	ALABAMA	AL	BLOUNT	1009	US PRESIDENT	RALPH NADER	GREEN	154	17973	20220315	TOTAL
2000	ALABAMA	AL	BLOUNT	1009	US PRESIDENT	OTHER	OTHER	175	17973	20220315	TOTAL
2000	ALABAMA	AL	BULLOCK	1011	US PRESIDENT	AL GORE	DEMOCRAT	3395	4904	20220315	TOTAL
2000	ALABAMA	AL	BULLOCK	1011	US PRESIDENT	GEORGE W. BUSH	REPUBLICAN	1433	4904	20220315	TOTAL
2000	ALABAMA	AL	BULLOCK	1011	US PRESIDENT	RALPH NADER	GREEN	24	4904	20220315	TOTAL
2000	ALABAMA	AL	BULLOCK	1011	US PRESIDENT	OTHER	OTHER	52	4904	20220315	TOTAL
2000	ALABAMA	AL	BUTLER	1013	US PRESIDENT	AL GORE	DEMOCRAT	3606	7803	20220315	TOTAL

year	fips	population	w_population	b_population	o_population	nh_population	hi_population	na_population	male_population	female_population	age0_population	age1_population
1990	1025	27289	15579	11643	35	27196	93	0	13052	14237	413	1564
1990	1031	40293	32869	6950	160	39831	462	0	19673	20620	582	2146
1990	1041	13598	10068	3516	11	13576	22	0	6421	7177	179	733
1990	1053	35526	24377	10050	1045	35378	148	0	17454	18072	484	1854
1990	1101	209537	119702	87856	415	207933	1604	0	98854	110683	3508	12717
1990	1115	50084	45376	4502	134	49894	190	0	25156	24928	764	2858
1990	1125	151095	109928	39625	252	150155	940	0	72887	78208	1993	7739
1990	2060	1388	916	1	466	1379	9	0	829	559	22	97
1990	2130	13956	11531	48	1913	13718	238	0	7315	6641	280	987
1990	2261	10024	8401	43	1265	9788	236	0	5506	4518	187	725
1990	2910	5705	1420	6	4262	5662	43	0	2996	2709	157	627
1990	4003	97918	89550	5196	868	69384	28534	0	49910	48008	1591	6185
1990	4021	116867	100023	3870	12396	82605	34262	0	60000	56867	2034	8048
1990	4025	108818	106188	321	1809	101793	7025	0	53268	55550	1259	5073
1990	5021	18096	18046	4	37	18036	60	0	8663	9433	206	846
1990	5069	85473	47874	37035	216	85056	417	0	41101	44372	1332	5078
1990	5073	9588	5898	3674	9	9566	22	0	4555	5033	142	511
1990	5079	13721	8793	4887	37	13599	122	0	8056	5665	164	670
1990	5099	10076	6899	3163	14	10022	54	0	4821	5255	137	518
1990	5107	28768	12953	15731	28	28548	220	0	13146	15622	574	2071
1990	5133	13752	12667	792	285	13133	619	0	6805	6947	183	714
1990	5145	54931	52871	1711	230	54570	361	0	26691	28240	725	2857

Data Input into Model

- Data from 30 bellwether counties (counties that have correctly predicted all or all but one presidential election outcome from 1980 through 2020)
- Engineered additional features: cumulative averages of voting metrics to enable model to be trained on historical data
  - Cumulative average voting ratios
  - Voter turnout ratio
  - Margin of victory (ratio-based)
  - Population growth ratio since last election year
- Feature Dimensions
  - Number of features: 32
  - Number of samples: 180
- Class Balance
  - 79 DEM wins
  - 101 REP wins
- Train-Test Split
  - 150 training samples: 2000 - 2016 presidential elections (30 x 5)
  - 30 testing samples: 2020 presidential elections (30 x 1)

Parameter Tuning & Results

- Parameter Tuning
  - Grid Search - runtime timed out :/
  - Random Search - best parameters:
    - n_estimators: 50
    - min_samples_split: 5
    - min_samples_leaf: 2
    - max_features: None
    - max_depth: 10
    - bootstrap: True
- Performance Metrics
  - Accuracy: 73.33% (22/30)
  - Precision:
    - DEM - 83.33% (5/6)
    - REP - 70.83% (17/24)
  - Recall:
    - DEM - 41.67% (5/12)
    - REP - 94.44% (17/18)