論　文

２０２２－２０２４米国大統領選の予測

CHEN WENHAO[†a]

Data Science Specialization (K1)(The Second Report)

あらまし　選挙予測に関する文献では，選挙における政党の勝利確率を予測する推論において，全体的な予測を捉えることができるアンサンブル手法が一番良い結果である。ここでは，古典的なロジスティック回帰との性能の違いを理解するために，最も一般的な分類モデルとアンサンブルモデルをいくつか検証する。ロジスティック回帰と様々な正則化ペナルティ項を用いて，2020 年と 2022 年の選挙予測に組み込まれた最も影響力のある特徴を探り，正則化の違いによって予測がどのように変わるかを理解する。正則化ロジスティック回帰モデルでは，世論調査データや州ごとの政党支持率など政治的要因に重きを置いているのに対し，集合モデルでは，人口統計学的特性など異なる特性を取り込む傾向があることがわかった。正則化は，モデルに新しい変数を追加する効果はなく，代わりに，標準的なロジスティック回帰と比較して，予測変数のセットの重要性を変化させるものであった。

キーワード　選挙予測，ロジスティック回帰，正則化，集合モデル

1. まえがき

本選挙予測におけるモデルは，ここ数十年で全米のメディアを一変させ，アメリカの選挙政治と選挙メディアシステムの定番となった。 選挙予測に関する最初の可能な議論は，1948 年に Louis H. Bean によって行われ[1]，将来の結果を予測するために、過去の選挙結果に焦点を当てたモデルが登場した。過去の大統領選挙の結果だけを取り上げたこの単純化されたモデルは，1940 年代から 1950 年代にかけて成功を収め，「嵐状態」などの政治学的概念を発展させる役割を果たしたのである。州内政党の勝利が連邦政党の勝利と相関することが多い州をストームステートと定義し，選挙予測における重要なトピックとなっている。

これらのモデルは，よりダイナミックなアプローチをとっており，多くの予測は標準的な分類のフレームワークだけでなく，アップデートされたベイズアプローチをとっています。その結果，アンサンブル手法、特にベイズ手法を取り入れた手法がこの領域では最も顕著である。ベイズ手法は，選挙結果に影響を与える最も有用な要因について「先験的」な仮説を立てない領域で最もよく機能することはよく知られている。多くの予測プラットフォームでは，様々な形式のアンサンブルが標準となっており，ランダムフォレストやブースティングなどの標準的なアンサンブル手法が一般的なアプローチの 1 つとなっています[2]。 おそらく最も一般的な予測方法は、反復回帰法を用いて何千もの予測を行い，最も可能性の高い結果のヒストグラムを作成することであろう。これにより，予測推論に寄与する確率分布の作成が可能となる。[3]

2. 背景

一番目の論文中で、アメリカ大統領選の揺れる州の重要性を確認した．また，いくつかの州は，毎年必ず共和党または民主党を選択し，残りのいくつかの州は，二つの党の間で揺れて，つまり，米国の大統領選挙、候補者の背後代表の党派の影響力は，彼らの箇人的な影響力よりも大きい。

そのため，この論文では、各州の投票データに影響を与える要因を，あらゆる角度から深く掘り下げて考えてみたい。

大統領選挙を予測するために 17 の人口統計的特徴と 16 の政治的特徴を混ぜ合わせ上院を予測するために 19 の人口統計的特徴と 20 の政治的特徴を混ぜ合わせましたこれらの特徴とその由来をまとめると次のようになる：

[†] 立命館大学情報理工学研究科，滋賀県
1-1-1 Nojihigashi, Creation Core #408
Kusatsu, Shiga, Japan. 525-8577
a) E-mail: gr0581vr@ed.ritsumei.ac.jp

・人口統計：州面積、州人口（2010、2020）、GDP（後の予測では一人当たり GDP を使用），教育達成率の割合（高卒%、大学卒%、専門学校卒%、学士号取得%、上級学位取得%），人種区分の割合（アメリカインディアン，アジア、黒人、ヒスパニック／ラテン，太平洋諸島民、白人），世帯年収中央値。 やや恣意的な選択ではあるが，上院の予測には州の雇用率や都市部の人口割合も含まれている。

出典：GDP/capita と都市部の割合以外の人口統計学的特徴は，すべて米国国勢調査の情報をもとに作成されている[4]。先ほどの後者 2 つの機能は，Wikipedia から引用した。

・政治：いずれの予測課題も，まず過去 4 回の大統領選挙の結果（2000/2004/2008/2012）、州の政党配置（PVI と呼ばれる指標による），現職大統領の地位（現候補が上院の議席/大統領職を獲得したかどうか），候補間の財政状態（選挙収入合計/支出合計），直近の大統領選挙の結果（2000/2004/2012）を含む。 直近の 5 つの世論調査の平均値。 上院予測には過去 6 回の選挙（すなわち，上院 2 議席をめぐる過去 3 回の選挙）の過去の実績を用い，大統領予測には過去 2 回の上院選挙を予測因子として用いた。 後の実験では，過去の大統領と上院議員の選挙結果を両予測から削除した。

出典：これらの政治的特徴の多くは，ウィキペディアから削り取ったものです。選挙資金に関する情報は Federal Election Commission[5]のデータから，世論調査の平均値は RealClearPolitics[6]の世論調査の平均値から手計算で求めたものである。特に，大統領選や上院選が無競争と予想される場合，一部の州では世論調査のデータがないため，これらの結果については，それぞれのレースにおける過去の選挙結果を推定値として使用した。地区 PVI は，Cook Political Report を出典としている[7]。現職有利は、数百のレースの 2018 年データに基づいている[8]。

どちらの予測も，それぞれのタスクに対して，異なる形式の学習とテストの分割を組み込んでいる。大統領選予測タスクでは，2016 年の大統領選結果とその世論調査/2010 年の人口数をタスクのトレーニングデータとして使用します。この場合の「テストセット」は 2020 年の特徴量であり，2020 年の大統領選挙を可能な限り忠実に予測することを目標とする。上院議員予測タスクについては，上記のように 2020-2022 年の特徴を全て保存する。過去の上院選を学習データとして用いるのではなく（1 回の選挙で上院 100 議席のうち 3 分の 1 しか争われないので難しい），2022 年サイクルに選挙がない州とその 2020 年の結果、および非競争州を上院選予測の学習データとして利用する。非競争州は、Wikipedia の 2022 年上院選挙に関するページで，予測結果がすべて「Solid」であるか「All-but-1」であるかで大まかに判断している[9]。このレポート執筆時点では、この基準に基づく競争州およびテストセットは以下の通り。アラスカ、アリゾナ、コロラド、フロリダ、ジョージア、ミズーリ、ネバダ、ニューハンプシャー、ノースカロライナ、オハイオ、ペンシルバニア。ワシントン州、ウィスコンシン州

3. 実験手法

本論文では，選挙予測モデルの成功に影響を与えるさまざまな特徴を探す。 ベイズ法は有効であるが，我々は探索空間を標準的なロジスティック回帰に現れる頻度論者の見解に留保している。 これは主に分類のためのロジスティック回帰のベースラインを保持するために行われますが，予測結果を改善するためにランダムフォレストとグラディエントブースティングによる予測結果を提供する。

最初の実験では，ロジスティック回帰、ラッソ正則化ロジスティック回帰、リッジ正則化ロジスティック回帰、決定木、ランダムフォレスト、勾配補強を使用した。 2 回目の実験では、これらのモデルのほとんどと，前述の Elastic-Net 正則化ロジスティックモデルを用い，α {0.2, 0.4, 0.6, 0.8}とした。 決定木は，1 回目の実験で学習がうまくいかなかったため，2回目の実験からは除外された。

上記の α パラメータに加え，このモデルで使用される他のハイパーパラメータは以下の通り：

・ロジスティック回帰モデル、正則化モデルのそれぞれに，ペナルティなしの標準モデルを含む標準スケールを適用した。 SAGA ソルバーによる各ロジスティックモデルの最適化。

・すべてのツリーベースのモデルにおいて，ランダム状態=83, RFs, GB は 120 の推定値を持つ。

表 1　最初の大統領予測実験のセットでは，予測において最も重要な特徴について報告している。
ロジスティック回帰モデルの係数推定値，および木ベースモデルにおける特徴の重要性が示されている。

Table 1: First set of experiments for presidential forecasting, most important features in prediction are reported.

Coefficient estimates for logistic regression models and feature importance for tree-based models are given.

Model	1st Important Feature	2nd Imp. Feat.	3rd Imp. Feat.	4th Imp. Feat.
Logistic Regression	State Party Align.	Lead Polling Party	3rd Recent Sen. Result	Median Age
	$(\hat{\beta}_1) = 1.960474$	$(\hat{\beta}_2) = 0.771840$	$(\hat{\beta}_3) = 0.689105$	$(\hat{\beta}_4) = 0.553478$
l1-Penalized Log. Reg.	State Party Align.	Lead Polling Party	3rd Recent Sen. Result	2012 Pres. Result
	$(\hat{\beta}_1) = 2.422856$	$(\hat{\beta}_2) = 0.655385$	$(\hat{\beta}_3) = 0.248870$	$(\hat{\beta}_4) = 0.217693$
l2-Penalized Log. Reg.	State Party Align.	Lead Polling Party	3rd Recent Sen. Result	2012 Pres. Result
	$(\hat{\beta}_1) = 1.239582$	$(\hat{\beta}_2) = 0.576008$	$(\hat{\beta}_3) = 0.4128377$	$(\hat{\beta}_4) = 0.409187$
Decision Tree	State Party Alignment	N/A	N/A	N/A
	$(\mathrm{FI}_1) = 1$	$(\mathrm{FI}_2) = 0$	$(\mathrm{FI}_3) = 0$	$(\mathrm{FI}_4) = 0$
Random Forests	State Party Alignment	Lead Polling Party	Median HH Income	% Adv. Degrees
	$(\mathrm{FI}_1) = 0.200451$	$(\mathrm{FI}_2) = 0.148936$	$(\mathrm{FI}_3) = 0.080751$	$(\mathrm{FI}_4) = 0.069273$
Gradient Boosting	State Party Alignment	2012 Pres. Result	% Adv. Degrees	Median HH Income
	$(\mathrm{FI}_1) = 0.267325$	$(\mathrm{FI}_2) = 0.170682$	$(\mathrm{FI}_3) = 0.096194$	$(\mathrm{FI}_4) = 0.095062$

表 2　表 1 と同じ，ただし上院の予測値。 注：このプロジェクトのポスターでは，特徴の重要度がほとんどのモデルで同じであると誤って報告されている。 これはコーディングミスによるもので，尖閣諸島予測モデルにおける正しい特徴量の重要度は以下の通りである。

Table 2: Same table as table 1 but for Senate forecasting. Note: The project poster for this project errroneously reported feature importance were the same for most of the models. This is not true, this was due to a coding error and the accurate, correct feature importance for the Senate forecast model are provided below.

Model	1st Important Feature	2nd Imp. Feat.	3rd Imp. Feat.	4th Imp. Feat.
Logistic Regression	Most Recent Sen. Result	2nd Recent Sen.	3rd Recent Sen.	% Urban Population
	$(\hat{\beta}_1) = 0.763433$	$(\hat{\beta}_2) = 0.727866$	$(\hat{\beta}_3) = 0.637550$	$(\hat{\beta}_4) = 0.509023$
l1-Penalized Log. Reg.	Most Recent Sen. Result	2nd Recent Sen.	3rd Recent Sen.	Lead Polling Party
	$(\hat{\beta}_1) = 1.14716$	$(\hat{\beta}_2) = 0.76192$	$(\hat{\beta}_3) = 0.358157$	$(\hat{\beta}_4) = 0.358157$
l2-Penalized Log. Reg.	Most Recent Sen. Result	2nd Recent Sen.	3rd Recent Sen.	Senate Incumbent Party
	$(\hat{\beta}_1) = 0.499051$	$(\hat{\beta}_2) = 0.470376$	$(\hat{\beta}_3) = 0.380422$	$(\hat{\beta}_4) = 0.354815$
Decision Tree	Most Recent Sen. Result	N/A	N/A	N/A
	$(\mathrm{FI}_1) = 1$	$(\mathrm{FI}_2) = 0$	$(\mathrm{FI}_3) = 0$	$(\mathrm{FI}_4) = 0$
Random Forests	Most Recent Sen. Result	2nd Recent Sen.	Lead Polling Party	2020 Pres. Result
	$(\mathrm{FI}_1) = 0.187645$	$(\mathrm{FI}_2) = 0.111745$	$(\mathrm{FI}_3) = 0.0875037$	$(\mathrm{FI}_4) = 0.079881$
Gradient Boosting	Lead Polling Party	2020 Pres. Result	% Adv. Degrees	Median HH Income
	$(\mathrm{FI}_1) = 0.238209$	$(\mathrm{FI}_2) = 0.164538$	$(\mathrm{FI}_3) = 0.127514$	$(\mathrm{FI}_4) = 0.125685$

・ブースティングでは，ベーススコアを 0.5、1 本の木に使用する特徴の割合＝0.3、学習率 = 0.1、L1 正則化のための α = 10 とした。

3．実験

前述のように，最初の実験セットでは，上記のモデルのセクションで説明したすべての機能とモデルを使用した。 2 つ目の実験では，同様の機能を使用したが，以下のように変更した：

・上記の通り，決定木モデルが削除され，elasticnet 正則化論理モデルが追加された

・州の GDP の特徴を一人当たり GDP の予測に変換したが，これは主に州がモデルにもっとニュアンスを加えるかどうかを評価するためであった。一人当たり GDP は，人口動態のニュアンスをあまり捉えられない州全体の GDP よりも，州の特性をよく捉えている。

・上院の予測では，特徴として，過去の 4 つの上院の結果と 2000/2004/2008 年の大統領選挙の結果を取り除いた。 大統領選の予測では，過去の上院選挙の結果を完全に削除した。 これは，特に上院の場合，予測に関連性の高い他の特性を理解するために行ったものである。

セグメンテーションとモデリングは，実験、トレーニング、テストを通じて不変である。

表 3　上院の予測に関する第二の実験セット，予測における最も重要な特徴を報告する。
Table 3: Second set of experiments for Senate forecasting, most important features in prediction are reported.

Model	1st Important Feature	2nd Imp. Feat.	3rd Imp. Feat.	4th Imp. Feat.
Logistic Regression	Recent Sen. Result	Incumb. Sen. Party	2020 Pres. Result	Lead Polling Party
l1-Penalized Log. Reg.	Recent Sen. Result	2020 Pres. Result	Lead Polling Party	Incumb. Sen. Party
E.Net LR ($\alpha = 0.8$)	Recent Sen. Result	Incumb. Sen. Party	2020 Pres. Result	Lead Polling Party
E.Net LR ($\alpha = 0.6$)	Recent Sen. Result	Incumb. Sen. Party	2020 Pres. Result	Lead Polling Party
E.Net LR ($\alpha = 0.4$)	Recent Sen. Result	2020 Pres. Result	Lead Polling Party	Incumb. Sen. Party
E.Net LR ($\alpha = 0.2$)	Recent Sen. Result	Incumb. Sen. Party	2020 Pres. Result	Lead Polling Party
l2-Penalized Log. Reg.	Recent Sen. Result	Incumb. Sen. Party	Lead Polling Party	2020 Pres. Result
Random Forests	Recent Sen. Result	% Adv. Degrees	2012 Pres. Result	Lead Polling Party
Gradient Boosting	Rec. Sen. Res. (Sen. 2)	Party Alignment	% Lead Polling Party	Recent Sen. Result

４．結果

実験１：最初の実験セットについては，表 1 および表 2 に，各モデルにおける最も重要な予測変数の要約結果を見ることができる。 大統領選挙の予測では，（PVI 指標に基づく）州の政党の配置が，州の投票方法の最大の予測要因の一つであることがわかる。 その他の重要な指標としては，世論調査でリードしている政党，過去の上院選挙の結果，さらにはバカロレア取得後の人口の割合など，ユニークな人口統計学的な指標もある。 これは，特に世論調査や政党の配置など，ほとんどの大統領選の予想と一致している。

上院の予測については，過去の上院選挙結果が 2022 年の予測の最大の指標であることは明らかである。 これは，標準的な予測文献で予想されることとは少し異なるので，2 番目の実験でこの特徴量の重要度の偏りを説明する。

予測に関しては，各大統領選挙は全く同じカテゴリーに分類され，2020 年以降の州はアリゾナ、ジョージア、ミシガン、ペンシルバニア、ウィスコンシンを除いて全て同じカテゴリーに分類されます。これらの州はすべて，2016 年の選挙と 2020 年の選挙で党派が逆転している。 上院の予想については，一部の「競合州」が同じ政党にまとめられている一方，ほとんどの州で予想に幅がある。 図 1 は、これらの結果をまとめたものである。

実験２：大統領選挙の予測では，ほとんどのモデルが世論調査の第一党，州の政党の整合性，2012 年と 2008 年の大統領選挙結果の一部の整合性を予測の政治的特性として使用している級学位を持つ人の割合と世帯収入の中央値は、最も重要な人口統計学的予測因子である。 各モデルは，世論調査における有力政党と政党の並びを各モデルの予測の最上位特徴として使用した。 以前のデータセットと結果が似ているため（予測値を含む），新しいデータセットにおけるこれらの予測セットに関するこれ以上の考察は省略します。

上院の予測については，過去の上院選挙がこれだけ特徴量セットから削除されているのだから，モデルの最重要特徴量に含まれる特徴量の種類に変化があるかどうかを観察してみる。 世論調査における政党のリード，現職上院議員の現職の状況（つまり、現在議席を持つ上院議員の所属政党），2020 年の大統領選挙の結果，最近の上院の結果，そしてモデルにおける主要予測因子がある程度一致していることが分かる。表 3 はこれらの結果をまとめたもので，ロジスティック回帰モデルはどれも似たような特性を持ち，正則化がある特性の重要度の順序に大きく影響していることがわかる。

スパース性については、LASSO と Elastic-Net の α ＞0 モデルは，全体的に特徴量が減少していることがわかる。LASSO では、23 の特徴量がゼロになり，上記の 4 つの特徴量と 2008 年の大統領選挙の結果のみが予測因子となった。

各モデルでは，勾配補強に加えて，各州内の 2/3 カテゴリの上院議員の直近の結果が、結果を予測する上で重要な特徴となっている。 注目すべきは、大統領選挙の結果を使うのではなく，ブースティングは州の上院議員の直近 2 回の選挙を予測の特徴として使っていることである。 また、これらの予測モデルには，人口統計学的特性はほとんど含まれず，政治的特性が支配的であることがわかる。 州の人口統計学的特性を重要視しているモデルは，上級学位保持者の割合を 2 番目に重要な特性としてモデル化した Random Forest

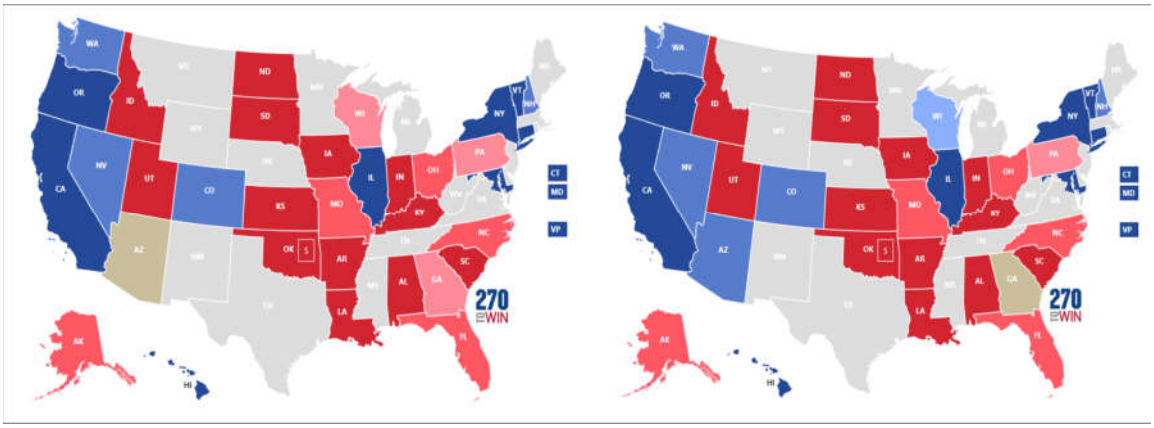

図 1　　左：第 1 回実験データセットによる上院の予測。右側。2 番目の実験のデータセットによる上院議員予測。どちらのビジュアルも，色は各実験の 6 または 8 モデルにおいて，ある州がどの程度強く特定の政党に向かうと予測されたかを示している。赤の実線と青の実線の州は，トレーニングの一部として使用された非競争州である。茶色は「純粋なトスアップ」であり，半分の予測は民主党の勝利を，半分は共和党の勝利を与えた。ピンクとブルーの色調の変化は，モデルのセットから得られた特定の予測が，どの程度強く特定の政党に勝利したかを反映しています。　水色やピンクの最も濃い色（青や赤でない色）は，8 つのモデルすべてで 1 つの政党しか勝利しないと予測された州です。注目すべきは，アリゾナ州、ジョージア州、ウィスコンシン州の予測差である。

Figure 1: Left: Senate forecast from the first experiment dataset. Right: Senate forecast using the second experiment's dataset. For both visuals, the color indicates how strongly a state was predicted towards a given party over the 6 or 8 models in each respective experiment. Solid red/solid blue states are noncompetitive states that were used as part of training. Brown is a "pure tossup" - half the predictions gave a Dem win for the state while half gave it to Republicans. Varying shades of pink and blue reflects how strongly certain predictions from the sets of models gave it to a certain party. For the darkest shade of light blue or pink (that aren't solid blue or red), these were states where every model predicted only one party winning across all eight models. Notable state differences in forecasts include Arizona, Georgia, and Wisconsin.

モデルのみである。

予測に関しては，大統領の結果は最初の実験とほぼ同じであり，これは新しいデータセットでは特徴の重要度にそれほど差がないため，理にかなっていると言えるでしょう。 上院の予測については，8 つのモデルの全予測の平均をとったところ，最初の実験での予測とはかなり異なる結果が得られました。 これらの結果の概要を図 1 に示す。

５．むすび

結果から，モデルは使用した政治的特徴とかなり一致しており，これらの特徴は現代の予測と最も一致していることが観察される。 政党支持率、現職の地位、州内の投票率は，選挙予測において最も強い指標であり，上記の特徴分析から，学習特徴を変更した後でも，これらの特徴が強く強調されていることがわかります。人口統計学的特性はほとんど見られないが，上級学位保有者の割合や世帯収入の中央値など，いくつかの特性は政党の勝利を強く予測するものである。 このことは，米国における長年の政界再編の傾向と一致している。裕福で教育水準の高い州は常に左傾化し，右傾化した州は大学教育水準が低く，しばしば経済的困難に直面する傾向がある。

正則化の効果については，2 つ目の実験セットでは，正則化によって新たな特徴がもたらされないことが確認された。 その代わり，ElasticNet α を変更すると、新しい機能を導入するのではなく，ある機能の特徴的な重要度が変わる。これは正則化に関する文献である程度予想されていることで，係数は変わるが，特徴構造は全体的に似たようなままである。 より明白な特徴の違いについては，統合されたアプローチは，上院の予測にはほとんど存在しない人口統計学的特徴を含む，予測力を持つ他の特徴についての洞察を提供することが示された。 上院モデルでは，勾配補強自体が上院の予測に新たな機能をもたらすことが示された。

もし，私がこのプロジェクトにもっと時間があれば，モデルのいくつかの重要な変更を検討したと思う。まず，ベイズモデリングをより深く探求し，最新の統合ベイズ法が，機械学習で顕著な頻度学習とテスト集合の分割を用いずに予測を行う方法を理解したいと思いる。 この先も，完全予測に貢献する機能最適化の研究を続けていくことを目標としている。 今後は、州の密度、医療アクセス統計、平均寿命などの人口統計学的特徴や，より詳細な世論調査や選挙資金データ

などを盛り込んでいく予定である。 これらの指標の多くは，最新のモデルで顕著に見られるものだが，多くの場合，慎重に選択された分析とチューニングが必要である。 最後に，大統領選挙における州の反転を予測しようとすると，困難な作業であることが証明されており，しばしば過去の選挙結果にオーバーフィットしてしまうため，州の結果自体を予測しようとするのではなく，州の選挙予測がいつ変化するかを予測するという別のアプローチも検討したい。

文　　献

[1] Bean, Louis Hyman. How to Predict Elections. Greenwood Press, 1972.

[2] Montgomery, Jacob M.. Improving Predictions Using Ensemble Bayesian Model. Political Analysis, 2012-03-22. https://www.jstor.org/stable/23260318

[3] natesilver538. “A User’s Guide to Fivethirtyeight’s 2016 General Election Forecast.” FiveThirtyEight, FiveThirtyEight, 29 June 2016

[4] https://data.census.gov/profile?q=United+States&g=0100000US

[5] https://www.fec.gov/data/

[6] https://www.realclearpolitics.com/epolls/latest_polls/senate/

[7] https://www.cookpolitical.com/cook-pvi/2022-partisan-voting-index/state-map-and-list

[8] https://fivethirtyeight.com/features/how-much-was-incumbency-worth-in-2018/

[9] https://en.wikipedia.org/wiki/2022_United_States_Senate_elections#Predictions