Who controls U.S. politics? An analysis of major political endorsements in U.S. midterm elections

Ziang Huang

1. Introduction

Former president Donald Trump endorsed over 200 candidates during the 2022 election cycle; in total, he has endorsed 551 candidates since he took office (Ballotpedia, 2022). Facilitated by the convenience of social media, other major political figures, including Bernie Sanders and Barack Obama, have endorsed similar numbers of candidates. Though such endorsements ostensibly advance a certain political cause and support a candidate's viability, it is uncertain what effects they actually have on the political landscape of the United States today. In this project, we want to answer two crucial questions: first, the reasons and underlying patterns behind the midterm endorsements of politicians in the US, and second, how much of a measurable effect these endorsements have on midterm elections in the United States and the political makeup of Congress as a whole. In turn, we hope to ascertain the degree to which major political figures such as Trump or Sanders still retain lingering influence over their party. Training our model on data from 2018 and 2020, we will finally generate probability predictions for each Senate election occurring in the 2022 midterms alongside the overall probability for each party to gain control of the Senate.

To clarify patterns and observable trends in what candidates each political figure endorses, we applied density clustering techniques to fundamental variables in our data set (district political leaning (Cook Political Report, 2017), district education level/income, race and age demographics), then draw comparisons with nation-wide averages in the US. We then performed a logistic regression on Sanders' and Trump's endorsement data, and discovered that endorsement behaviors fell into one of two categories: pragmatic and ideological. Trump's 2022 endorsements were 85% predictable through data on his 2018 and 2020 endorsements' demographic variables alone, with PVI (political leaning) contributing 55% towards the endorsement status of a candidate, demonstrating Trump's preference for solidly Republican states compared to ideologically aligned candidates. On the other hand, Sanders' endorsements were only 67% predictable with fundamental variables and 80% predictable through the progressivity of candidates, showing an inclination to endorse based on individual political leanings.

In order to quantify the effect of an endorsement on an election, we applied a binary support vector machine classification model on 2018 and 2020 midterm results with the endorsement of a specific politician as a variable, alongside fundamental demographic factors such as race, income and age distribution in states. It was discovered that Trump's and Sanders' endorsements were a good predictor for midterm winners, but did not significantly contribute to the election result (11% and 7% respectively), with main contributors instead being Cook's PVI (political leaning of state) and education level. Using the 2018 and 2020 data sets, our model generated probabilities for a Republican victory in each state in the 2022 Senate elections and predicted a 67% chance for Democrats to retain the Senate in the 2022 midterms, with endorsements by Sanders, Trump and others having no meaningful impact.

2. Who, Where and Why do Candidates Endorse?

This section aims to answer one question: who, where, and why do major political figures endorse in US midterm elections? What are common trends in the endorsements of each figure in terms of state demographics (race, age, income attainment and political leanings), especially when compared with nation-wide data?

2.1 Methodology

The fundamental variables we consider are related to the district the endorsed candidate hails from, and includes: support of endorsing figure in district/state, Cook's PVI for district (Cook's Political Report, 2017), median income of district, percentage with Bachelor's Degree in district, percentage white in district, and percentage aged 65 or over in district (US Census Bureau, 2021). To identify these trends, our methodology involves first separating the six variables into three groups of 2 according to their cross-correlations for dimensionality reduction, then clustering the resulting two-dimensional data through a density-based DBSCAN method (Ester et al., 1996). We will calibrate the parameters of DBSCAN through a k-Nearest Neighbor algorithm to find the distance between points in the data set. Finally, to summarize factors affecting which candidates will receive endorsements, we will use a binary-classification logistic regression model outputting one of two states (1: endorsed, 0: not endorsed) from the fundamental variables stated above.

DBSCAN relies on two parameters: ϵ, the minimum distance between two points required for them to be considered reachable from each other, and $minPt$, the minimum number of points required to define a cluster (Ester et al., 1996). If a point P is within distance ϵ of point Q, then Q is considered *directly reachable* from P; if Q can be reached from P through the sequence $\left(P, P_1\right), \left(P_1, P_2\right), \cdot, \left(P_n, Q\right)$ where each pair of points $\left(P_i, P_{i+1}\right)$ are directly reachable from each other, then Q is *reachable* from P. A cluster with core point P is then defined as a set of points, at minimum $minPt$ points, that are reachable from a point P.

A k-Nearest Neighbor (k-NN) algorithm is used to find the optimal value of ϵ in DBSCAN (Sharma and Sharma, 2017). Define the set of nearest k neighbors to a point X as $S_k(X)$ such that $\left|S_k(X)\right| = k$ and belongs to the data set. For every $A' \notin S_k(X)$:

$$dist(X, A') \geq \max_{A \in S_k(X)} dist(X, A)$$

where $dist(X, A)$ is the Euclidean distance between X and A. Thus, $S_k(X)$ contains the k nearest points to point X. The average distance $\mu_k(X)$ is the mean of all $dist(X, A)$ for $A \in S_k(X)$. $\mu_k(X)$ for all points $X \in$ data set D is then sorted in ascending order and plotted, and the value for ϵ applied in DBSCAN clustering is the value of $\mu_k(X)$ at the "knee" of the curve, found through the knee() method in Python. k will be set to 5 for our model. The value of $minPt$, as is usual in two-dimensional data, is set to 4.

Finally, we use a binary-classification logistic regression model based on the previously stated fundamental variables with an additional variable of incumbency to determine the possibility of a candidate receiving an endorsement (Ballard et al., 2020). The regression equation is based on the sigmoid function $\frac{1}{1+e^{-x}}$, and comes in the form

$$\frac{e^{\sum_{i=1}^{n} a_i x_i}}{1+e^{\sum_{i=1}^{n} a_i x_i}}$$

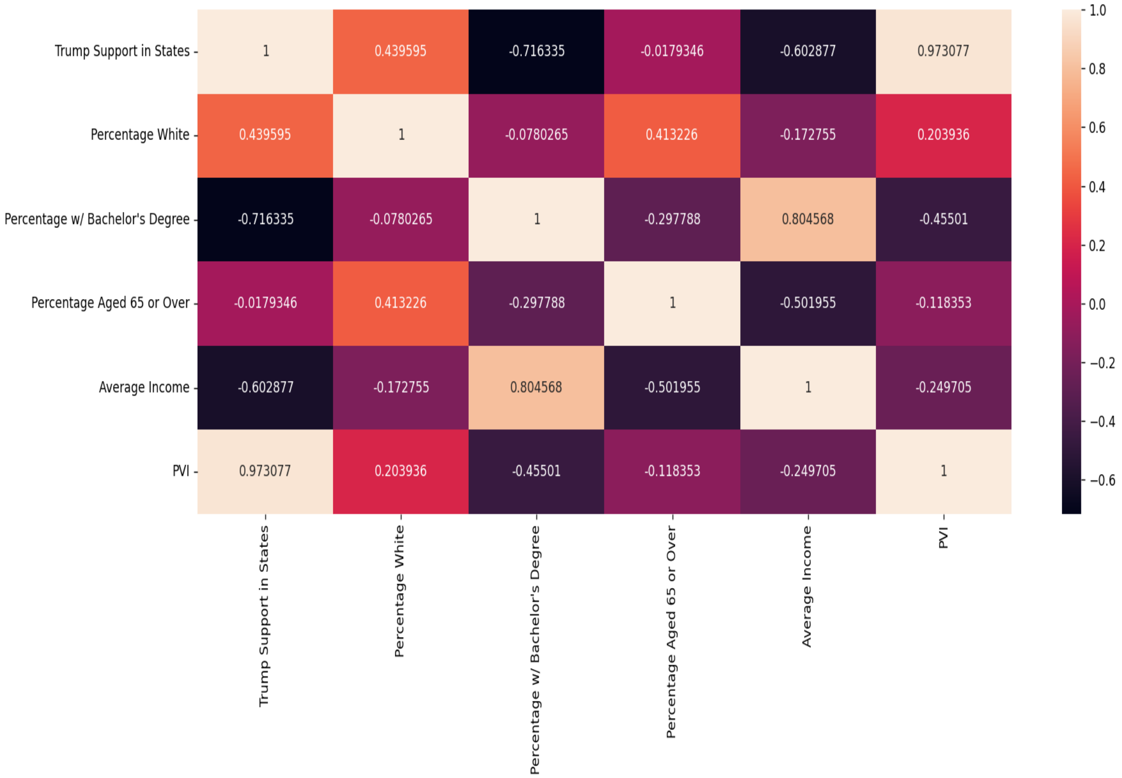
where $x_1, ..., x_n$ and $a_1, ..., a_n$ are the n independent variables and coefficients assigned to each variable respectively. We will also validate and evaluate the logistic regression through a confusion matrix metric:

$$\begin{gathered}[[TN,\ FN]\\ [TP,\ FP]]\end{gathered}$$

where TP, FP represent correctly/falsely predicted positive values and TN, FN represent correctly/falsely predicted negative values respectively. Through the confusion matrix, we will assign an accuracy score $\frac{TP+TN}{TP+FP+FN+TN}$, the proportion of true predictions to total data points.

2.2 Donald Trump's Endorsements

Former President Donald Trump is perhaps the most prolific endorser among all US politicians both during and after his presidency, endorsing a total of 556 candidates (governor, Senate and House) through a combination of social media and appearances at campaign rallies (Ballotpedia, 2022). Drawing on his US House and Senate endorsements in the 2018 midterms, we obtain the following correlation matrix between the fundamental variables:



	Trump Support in States	Percentage White	Percentage w/ Bachelor's Degree	Percentage Aged 65 or Over	Average Income	PVI
Trump Support in States	1	0.439595	-0.716335	-0.0179346	-0.602877	0.973077
Percentage White	0.439595	1	-0.0780265	0.413226	-0.172755	0.203936
Percentage w/ Bachelor's Degree	-0.716335	-0.0780265	1	-0.297788	0.804568	-0.45501
Percentage Aged 65 or Over	-0.0179346	0.413226	-0.297788	1	-0.501955	-0.118353
Average Income	-0.602877	-0.172755	0.804568	-0.501955	1	-0.249705
PVI	0.973077	0.203936	-0.45501	-0.118353	-0.249705	1

Fig 1. Correlation Matrix Between Fundamental Variables for Trump Endorsements

PVI and Trump support both represent political leaning of the district, and are highly correlated ($r = 0.959$); percentage w/ Bachelor's degree and median income both represent socioeconomic status, and are also highly correlated ($r = 0.823$); finally, percentage white and percentage 65 or over represent race/age demographics, and are most correlated with each other compared to any other variable, with $r = 0.525$. We will perform DBSCAN clustering on two-dimensional data based on these three groups.

2.2.1 Political Leaning

We will consider Trump support and PVI for all Trump-endorsed candidates to quantify trends in Trump-endorsed districts' political leanings. Implementing k-NN on the data set gives $\epsilon = 6$:

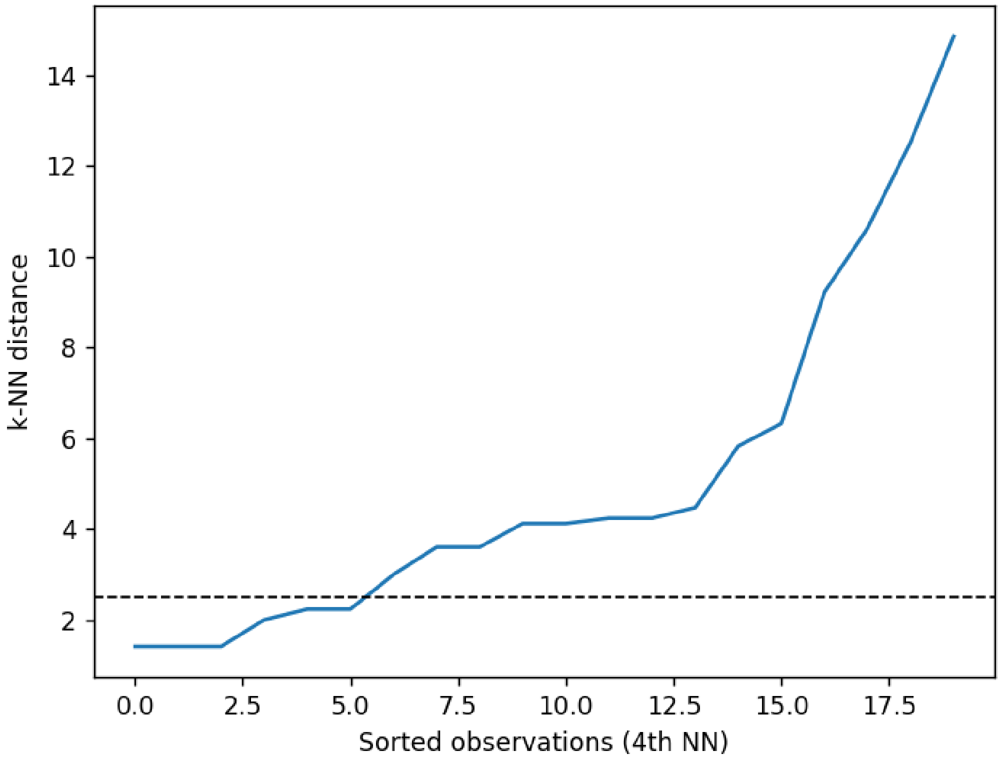


Fig 2. K-NN Plot for PVI vs Trump Support

DBSCAN clustering identifies the following cluster in the data:

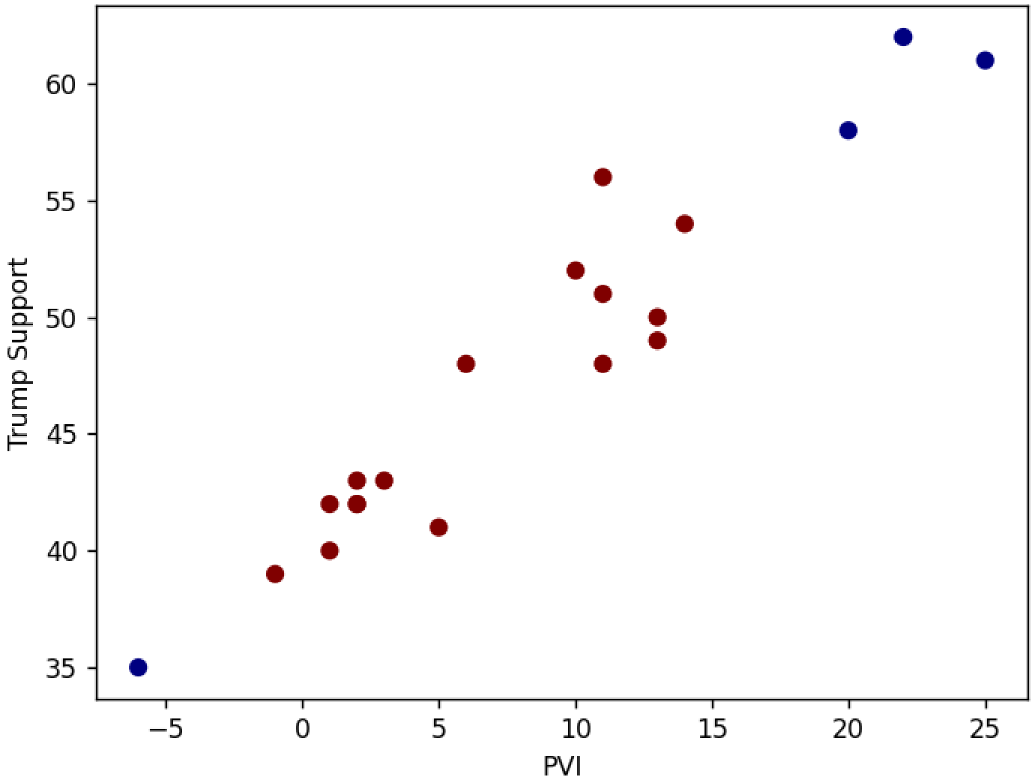


Fig 3. Clustered Political Leaning Data for Trump Endorsements

Comparing the primary cluster to nationwide PVI and Trump support figures, the average PVI of a Trump-endorsed state is +6.5 republican lean with standard deviation $\sigma = 5.20$ as opposed to a +3.56 nationwide average with $\sigma = 10.7$; Trump support for endorsed states also follows a similar trend, with an average support of 46.3% and $\sigma = 5.37$ as opposed to 44.0% and $\sigma = 9.12$ nationwide. We can therefore conclude that Trump tends to endorse, on average, states with 3% higher PVIs (republican leans) and 2.3% higher support for him, with the subset of states he places endorsements in being more concentrated over a smaller range of PVIs than the entire country.

2.2.2 Socioeconomic Status

The group of variables quantifying socioeconomic status of states/districts with endorsements contain median income and percentage with Bachelor’s degree (educational attainment). The k-NN algorithm obtains $\epsilon = 2500$. DBSCAN clustering obtains the following:

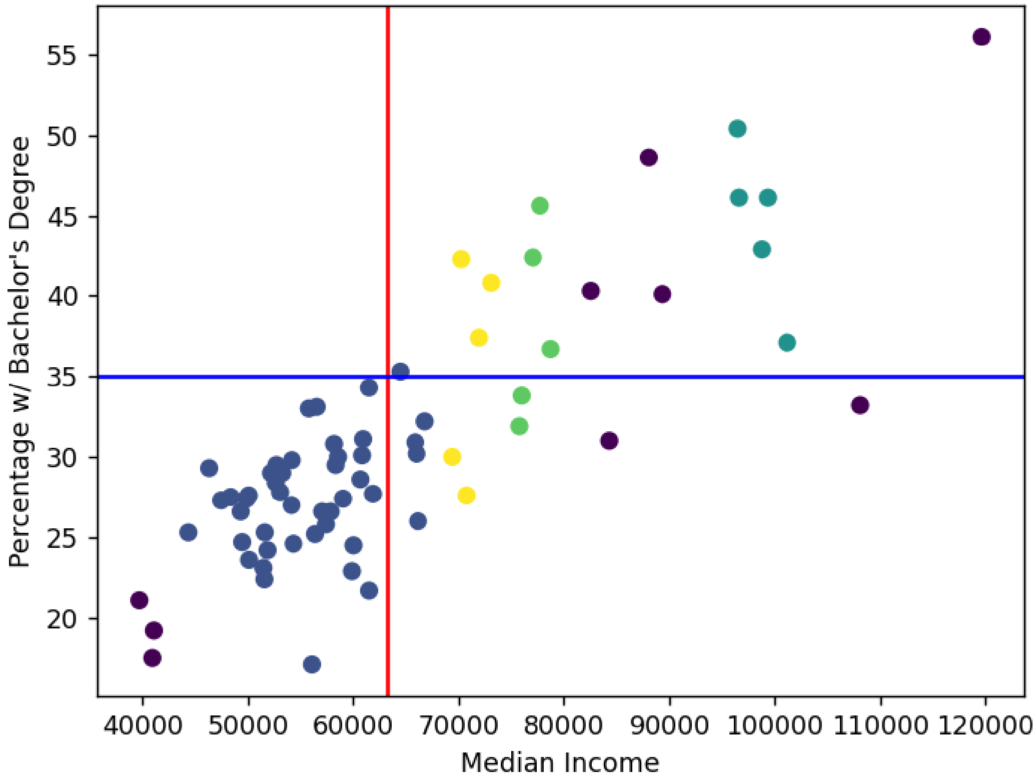


Fig 4. Density-Clustered Income and Educational Attainment for Trump Endorsements

The points in light blue represent the primary cluster with highest density, while the red vertical line represents the median household income of the US in 2018 (63,179$) and the blue horizontal line represents the percentage of the total US population with a Bachelor's degree. It is clearly observable that nearly the entire primary cluster of districts with the highest concentration of Trump endorsements fall below national income and educational attainment averages; this is true for the entire data set with educational attainment (31.0% with Bachelor's degrees vs. 35% nationwide), but median incomes between the data set and nationwide averages are comparable (64379 vs. 63179$) due to particularly rich districts such as New Jersey's 11th and New York's 2nd. This is more clearly demonstrated on a chloropleth map of states with Trump-endorsed candidates, colored according to Bachelor's degree attainment:

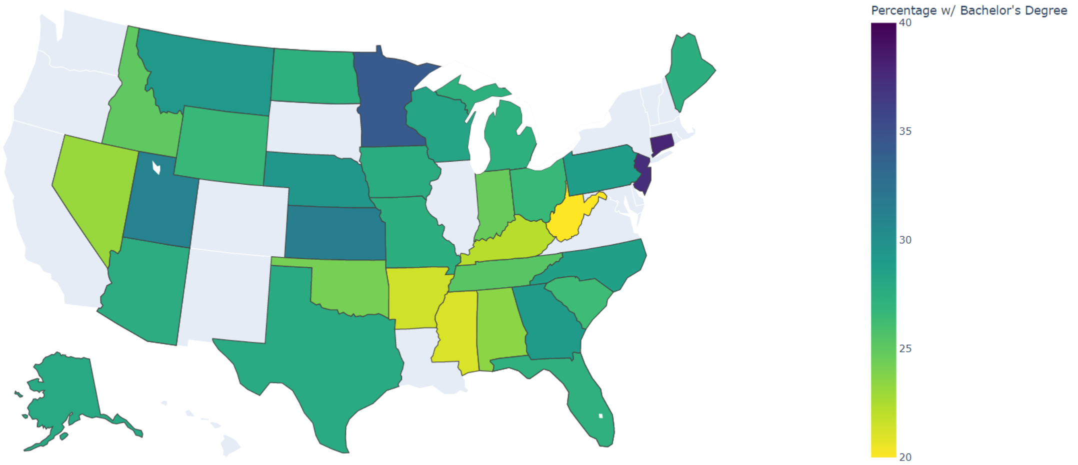


Fig 5. Chloropleth Map of Bachelor's Degree Attainment of Trump-Endorsed States

As exemplified on the map, most Trump-endorsed states fall between yellow (20%) and blue-green (30%), far below the national average of 35% (blue).

2.2.3 Age and Race Demographics

Age and race distribution comprise the remaining fundamental demographic variables for endorsed districts; the two variables are intercorrelated to a certain degree, with $r = 0.525$, allowing them to be grouped together for clustering. From k-NN, we obtain $\epsilon = 4$; DBSCAN clustering is as follows, with magenta representing the primary cluster of highest density, the horizontal blue line representing the percentage aged 65 or over throughout the US in 2018 (16%), and the vertical red line representing the percentage of whites in the US in 2018 (60%):

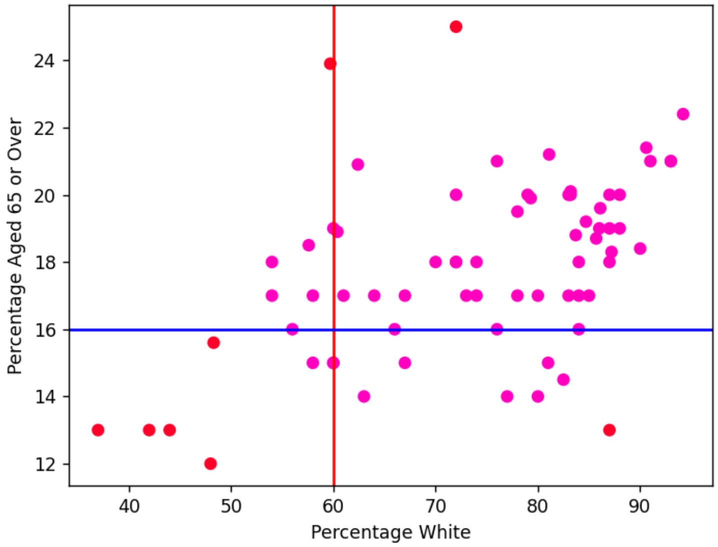


Fig 6. Density-Clustered Race and Age Demographics for Trump Endorsements

A trend of Trump's endorsed districts being both more predominantly white (ranging from 53% to nearly 100%, with highest concentration around 80 to 90%, as opposed to an average of 60% across the US) and older (ranging from 14% to 22%, with highest concentration around 17 to 20%, as opposed to an average of 16%). Data set averages confirm this, with Trump's endorsed districts being 73.8% white on average (vs. 60% countrywide) and 17.8% older than 65 (vs. 16% countrywide).

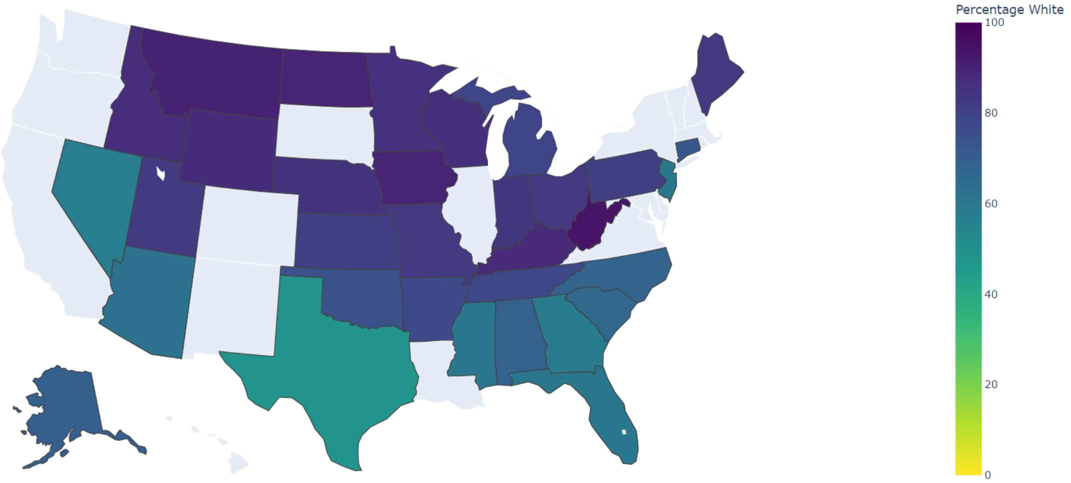


Fig 7. Chloropleth Map for Race Distribution of Trump-Endorsed States

The map above provides a clearer visualization: while the national average rests at 60% white, represented on the scale as a dark blue-green, Trump-endorsed states are predominantly light- to deep-blue (80 to 100% white).

2.2.4 Who Does Trump Endorse?

Applying logistic regression to our data set with an additional fundamental variable (incumbency of endorsed candidate), we obtain the following equation:

$$P(endorsement) = \frac{1}{1+e^{\sum_{i=1}^{6} a_i x_i}}$$

with coefficients $a_1,..., a_6$ corresponding to PVI, percentage white, incumbency, percentage with Bachelor's, percentage aged 65 or over, and median income:

Variable	Coefficient/Weight	Percentage Weight (Absolute)
PVI	0.925	55.0%
Percentage White	-0.00942	0.560%
Incumbent?	0.0221	1.31%
Percentage w/ Bachelor's Degree	-0.474	28.2%

Percentage Aged 65 or Over	-0.248	14.7%
Median Income	0.000398	0.0237%

This implies that the main contributing factor towards whether or not a candidate receives an endorsement from Trump is the political leaning of the state, measured by its PVI (55.0%), with the education level (28.2%) and the age distribution (14.7%) both having negative effects, while the other variables have a negligible impact. This provides a partial explanation towards none of the non-Trump-endorsed Senate candidates winning their elections in the 2018 midterms; Trump has a higher propensity to endorse candidates which are poised to win regardless, and vice versa. A glance at the political leaning of Trump-endorsed states, measured on a blue-to-red color scale according to PVI, reveals the extent of Trump's pragmatism:

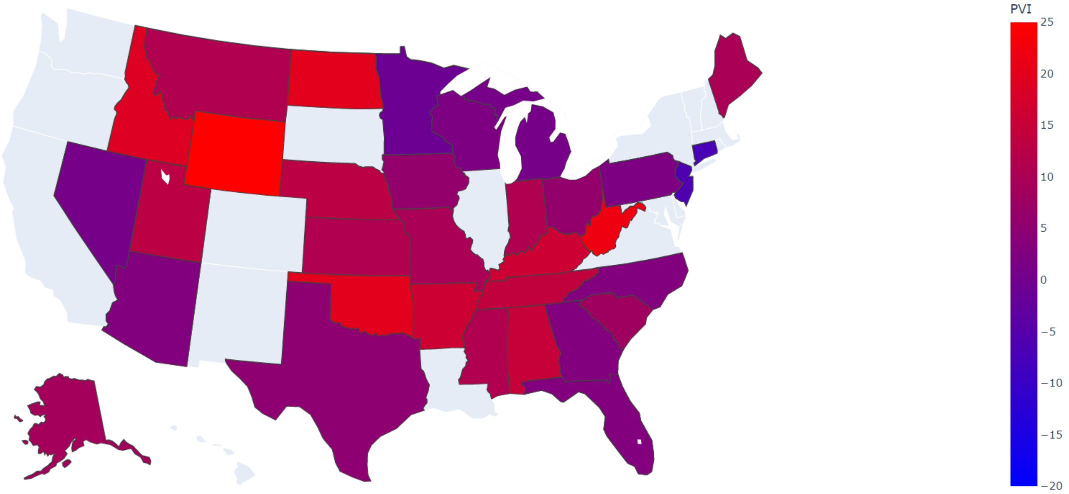


Fig 8. Chloropleth Map for PVI of Trump-Endorsed States

The confusion matrix obtained from cross-validation of the model is as follows:

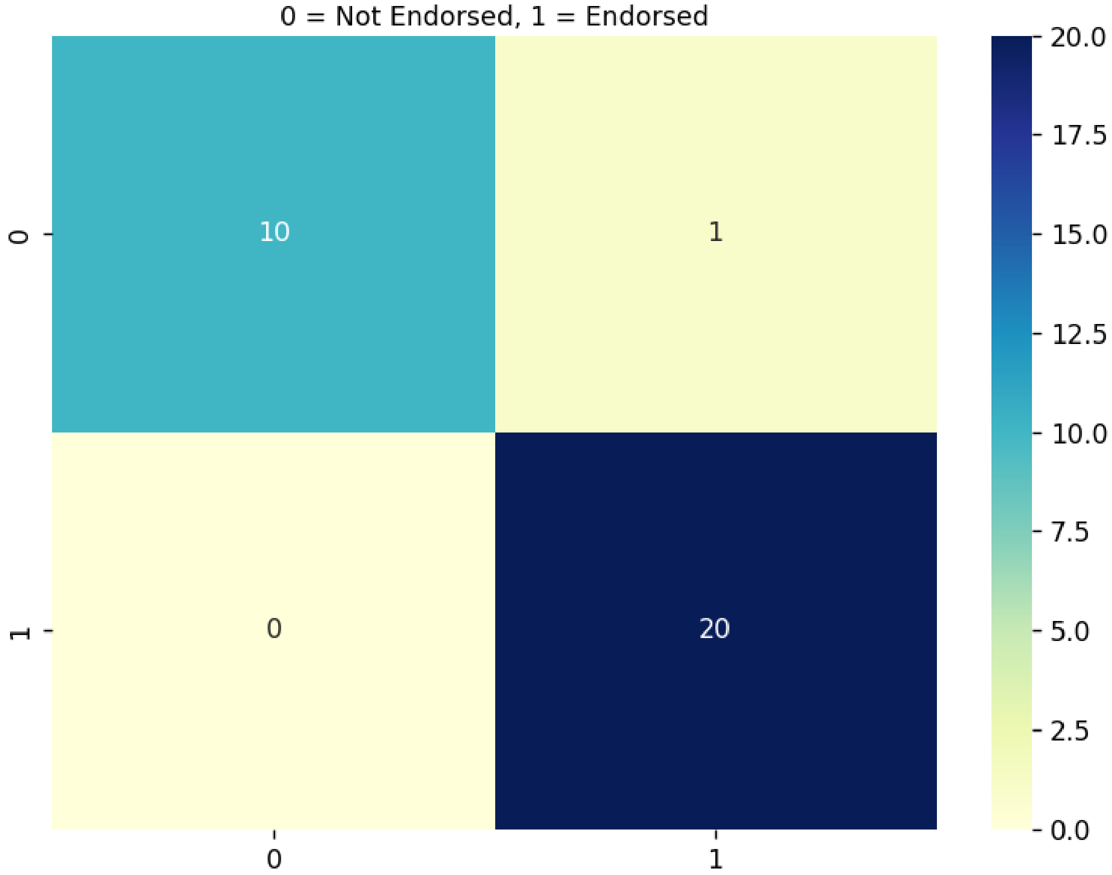


Fig 9. Confusion Matrix for Binary Logistic Regression of Trump Endorsements

which demonstrates a total of 10 true negatives to 1 false negative and a total of 20 true positives to 0 false positives at 97% accuracy. We further validated this by applying unseen 2020 midterm Senate election endorsement data into our model, obtaining an 85% accuracy level out of 34 candidates.

2.3 Bernie Sanders' Endorsements

Bernie Sanders is deservedly the foremost icon of progressive idealist politics in the United States today. A two-time candidate for President, Sanders alongside his previously marginalized platform of leftist social-democratic policies now enjoy national renown, with the next generation promising an influx of left-leaning progressives into Congress (Ocasio-Cortez, Omar, and many others). As Sanders continues to spearhead the progressive movement, he wields his continued influence over the Democratic Party in a wholly distinct way from Trump, as is most well-demonstrated through his endorsement patterns. As is expected, the fundamental difference between Sanders' endorsements and Trump's endorsements is Sanders' propensity to endorse only progressive left-leaning candidates regardless of state demographics or victory chances; in contrast, Trump's pragmatism in deciding to only endorse candidates in crucial swing states or in securely Republican states to "pad his record" reflects a lower level of alignment with any political ideology. We will visualize such trends through Congressional district-based and state-based chloropleth maps on Sanders-endorsed districts' political, racial and economic leanings.

2.3.1 Fundamental Trends in Sanders' Endorsements

From the lens of fundamental demographic variables - race, education level, income or political leaning - we were unable to discover any clear underlying trend in states Sanders endorses in. Beginning with political leaning, Sanders' Congressional endorsements do exhibit similar trends of tending to endorse Democrat-leaning states (average PVI -6.58 across 2018 and 2020 compared to Trump's average of +6.5), but closer examination reveals that the average only skews towards a Democratic PVI due to several firmly Democratic districts (California's 12th with -40 PVI, Georgia's 5th with -32 etc.), with Sanders not demonstrating the same reluctance to endorse candidates in politically disadvantageous districts as Trump did (West Virginia's 2nd district with PVI +22, etc.) The chloropleth map shows the wider distribution of political leanings across Sanders-endorsed states compared to Trump-endorsed states:

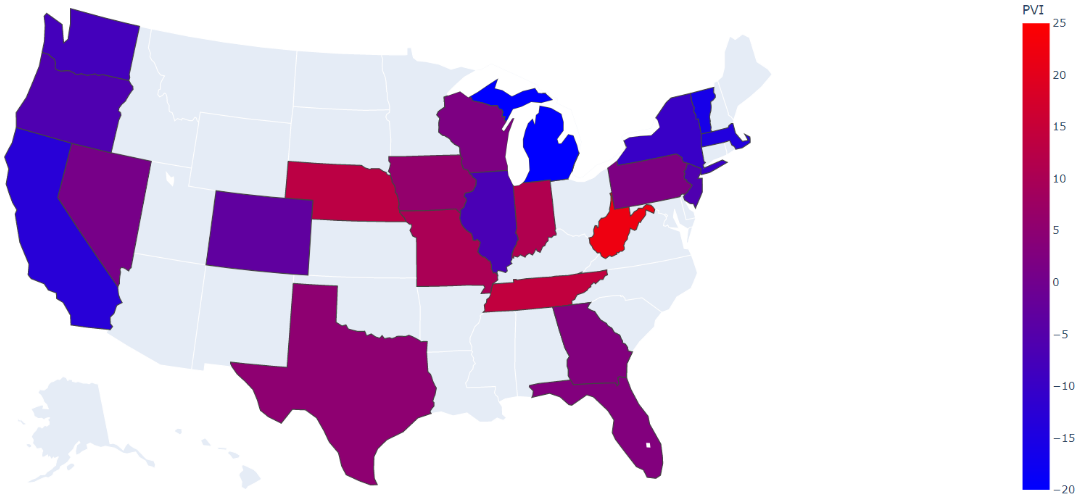


Fig 10. Chloropleth Map for Bernie-Endorsed States' PVI

The statewide map for PVI ranges across a wide spectrum from PVI = -20 to 22, with the presence of both deep-blue and deep-red states alongside a large amount of neutral purple states, contrasting directly with Trump's PVI map consisting of a purple to deep-red spectrum.

Both other demographic factors - educational attainment and race distribution - reveal no significant trends or major differences with national averages. Educational attainment, quantified by percentage attaining a Bachelor's degree in Sanders-endorsed states, is an average of 35.5%, consistent with the national average of 35%; the chloropleth map indicates a wide range of values from 20 to 40% instead of a consistent below-average distribution exhibited by Trump's educational attainment map.

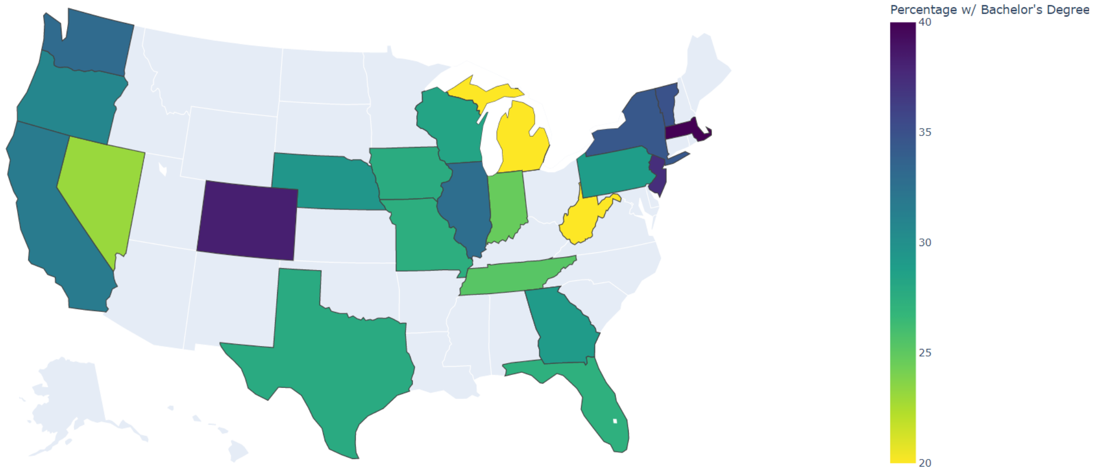


Fig 11. Chloropleth Map for Bernie-Endorsed States' Educational Attainment

Race distribution, measured by proportion of white residents, is an average of 63% across Sanders-endorsed states, barely varying from the national average of 60%, whereas Trump's endorsed states were heavily skewed towards high percentages of white residents (average 73.8%).

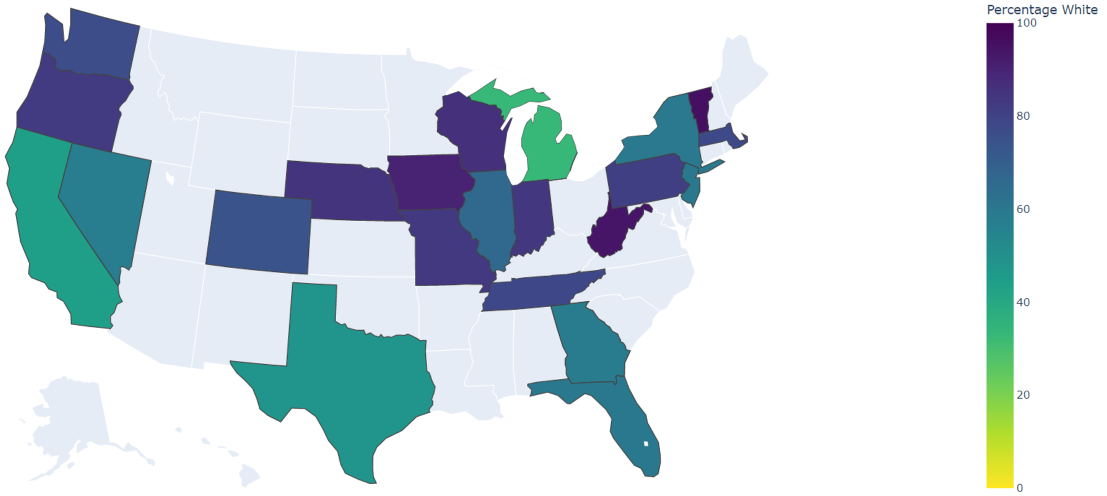


Fig 12. Chloropleth Map for Bernie-Endorsed States' Race Distribution

It is therefore unsurprising that the above fundamental demographic variables were poor predictors of Sanders' endorsements, achieving a relatively poor 67% accuracy:

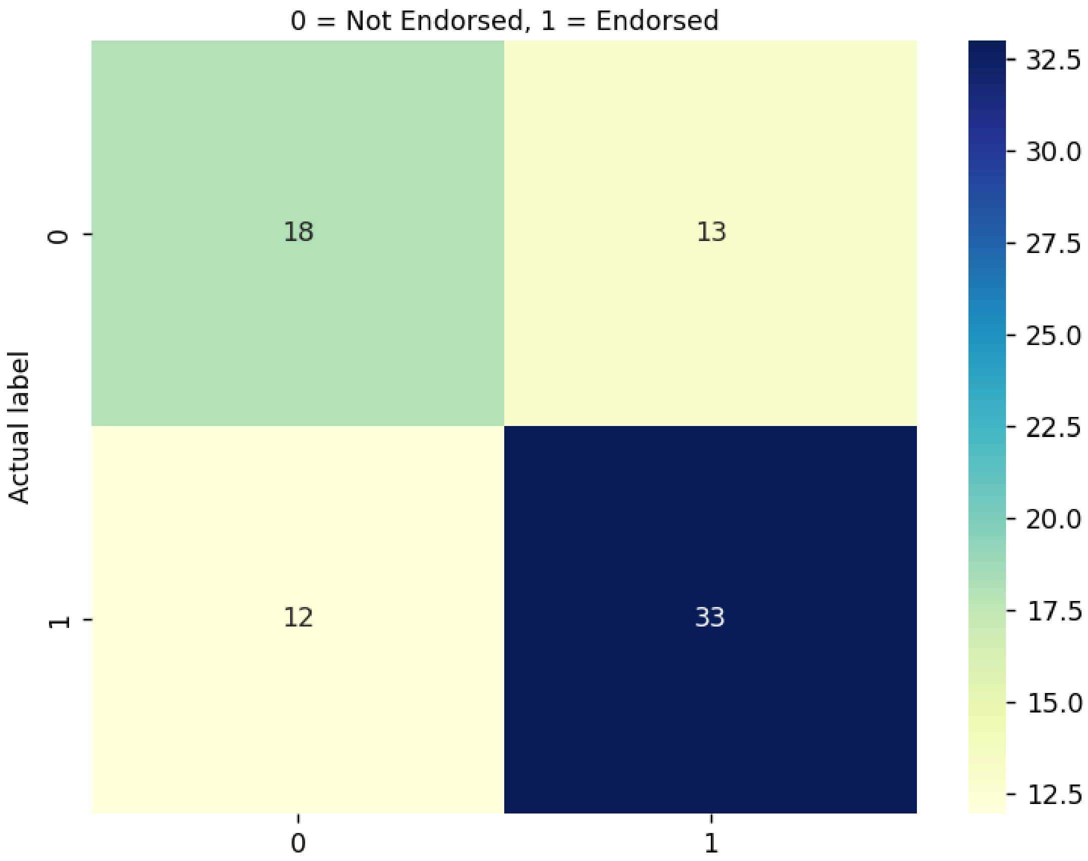


Fig 13. Confusion Matrix for Fundamental Model of Sanders' Endorsements

A total of 12 false negatives were recorded, to 18 true negatives (40% failure rate); 13 false positives occurred to 33 true positives (28% failure), underscoring the ineffectiveness of fundamental variables alone in predicting Sanders endorsements.

2.3.2 Pragmatism vs. Ideology: Understanding the Difference Between Sanders' and Trump's Endorsements

In the absence of reliable fundamental variables, we turn to the individual political ideologies of each candidate Sanders endorses. Using ProgressivePunch's progressivity scores for incumbent members of Congress (measuring the percentage of times a Congressperson or Senator votes in line with a "progressive" policy position), we attempted to identify whether or not the political ideology of incumbent Congresspeople and Senators would serve as a more reliable predictor of Sanders endorsements by using only the progressivity score as a variable in our logistic regression model. For non-incumbent candidates, we assumed that their progressivity scores were equal to the national average for Democratic Congresspeople and Senators (85% progressive). Progressivity alone was able to predict whether or not a candidate would be endorsed 80% of the time, with the dividing line being 88.34% (candidates above this figure would be classified as endorsed, and vice versa).

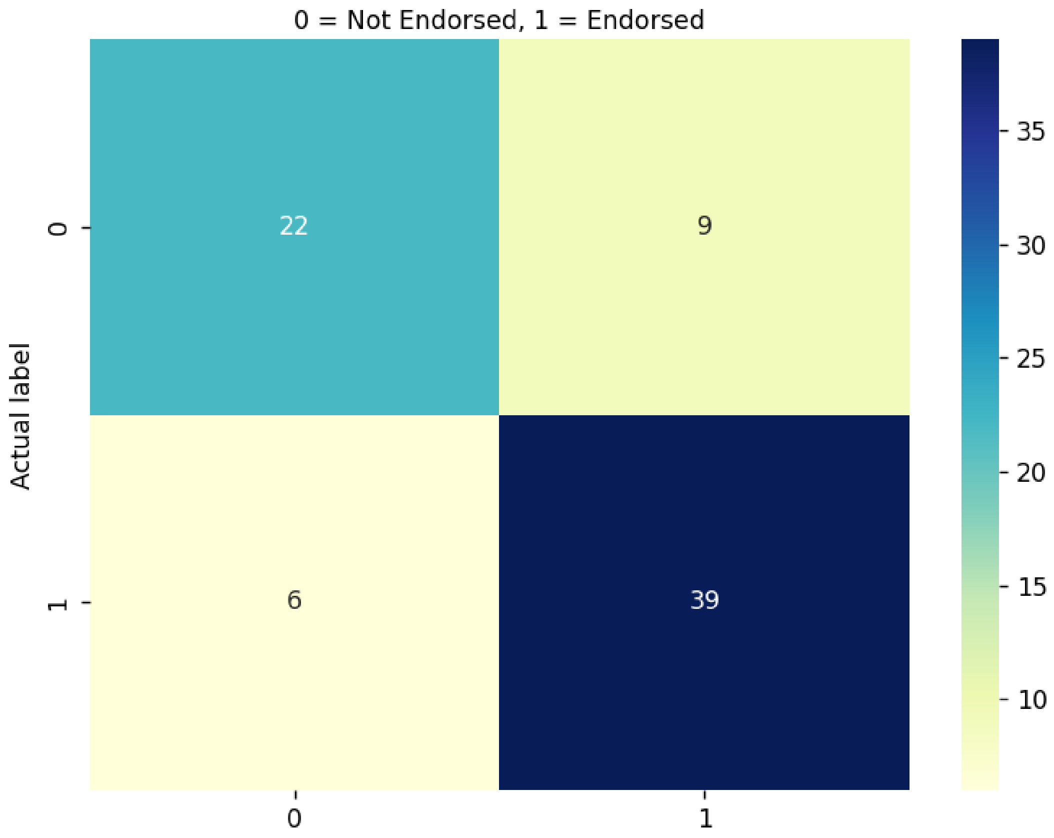


Fig 14: Confusion Matrix for Ideological Model of Sanders' Endorsements

The strongest conclusion we can draw from the examples of Donald Trump and Bernie Sanders is that endorsements by major US political figures can largely be divided into one of two categories: pragmatic and ideological. Pragmatic endorsements, using Trump as an example, are primarily influenced by factors outside the control of the endorsed candidate themselves: the state's political leaning, the education level of the state, and the racial distribution of the state - in short, whether or not the demographics of the state or district is favorable to the endorser's Party (e.g. Trump and the Republican Party). Indeed, Trump's endorsements are easily and accurately predictable through the fundamental variables outlined above. Ideological endorsements, on the other hand, are endorsements characterized and primarily influenced by whether or not the endorser supports the political ideology of the endorsed candidate themselves, and are largely better predicted by individual candidates' political positions and ideologies rather than fundamental variables.

3. The Effect of Political Endorsements

This section seeks to fulfill two goals: one, to measure the effects of political endorsements on election outcomes through probabilistic modelling, and two, to construct a predictive model based on endorsements and other fundamental factors using previously obtained results.

3.1 Methodology

We begin this section with an exploration of the methodology used in quantifying and predicting the power of endorsements. We model binary victory states as our independent variable (either 1 when a candidate wins, or 0 when a candidate loses), and introduce a new binary variable of endorsement (1 if endorsed by figure and 0 if not endorsed). Representing each candidate as a point in hyperspace, we use a linear support vector machine model (SVM) to conduct binary classification (Cotes and Vapnik, 1995). Let x be the seven-dimensional input vector of independent variables (including endorsement status, PVI, percentage white, incumbency, percentage with Bachelor's, percentage aged 65 or over, and median income). To ensure that the variables are similar in scale, we use the StandardScaler() function in Python's sklearn with $z = \frac{x-\mu}{\sigma}$, mean $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$, and standard deviation $\sigma = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \mu\right)^2}{N}}$. The linear SVM generates a hyperplane $H = w^T \cdot x + b$ that separates the positive data points (candidate wins) from the negative data points (candidate loses) based on a weight vector w^T normal to the plane and a bias/intercept b. As $w^T \cdot x$ is the projection of input vector x in the normal direction of the plane, the distance $d_H(x)$ between x and H can be calculated as follows:

$$d_H(x) = \frac{\left|w^T \cdot x + b\right|}{|w|^2}$$

To ensure that the hyperplane optimally separates the two data sets, we find the weight vector with arguments that maximize the closest distance between the positive/negative data with the hyperplane:

$$w^* = arg\ \max_w\ \min_n\ d_H\left(x_n\right) = arg\ \max_w\ \min_n\ \frac{\left|w^T \cdot x_n + b\right|}{|w|^2}$$

for optimal weight vector w^* and optimal hyperplane $H = w^{*T} \cdot x + b$. However, this assumes perfect separation of the two data sets; in order to account for scenarios where positives are sorted to the negative side and vice versa, we introduce an error variable ξ_n for an erroneously predicted point x_n (e.g. false positive):

$$\xi_n = C \cdot d_H\left(x_n\right)$$

directly proportional to the distance between x_n and the hyperplane, where C is an adjustable constant. Thus, we will attempt to optimize weight vector $w\ *$ as follows:

$$w^* = arg\ \max_w\ \min_n\ \frac{\left|w^T \cdot x_n + b\right|}{|w|^2} + \sum_{i=1}^{n} \xi_i$$

which represents the sum of the error terms and the previous optimization metric of minimum distance to H. This is achieved through the sklearn SVM module in python, and relies on Lagrangian multipliers. Ultimately, our classification for new data points is:

$$sign\left(w^T \cdot x_n + b\right) = +\ 1, candidate\ wins\ election$$
$$sign\left(w^T \cdot x_n + b\right) = -\ 1, candidate\ loses\ election$$

Our metric for the relative importance of each fundamental variable, obtained from the weight vector, is simply $I = \left|w^*\right|$, where the relative importance of each variable is measured by the absolute value of its corresponding coefficient in the weight vector (Guyon and Elisseeff, 2003). We deem a variable to negatively contribute to the election result if its weight is negative, and vice versa.

3.2 How Powerful Are Trump Endorsements?

Our model in this section will draw on data from Trump endorsements in both the 2018 and 2020 midterms. A total of 31 candidates were the Republican contenders for contested Senate seats in the 2018 midterms, with 19 receiving Trump endorsements and 12 not receiving an endorsement. Of the 19 endorsed candidates, 10 won and 9 lost the general election; of the 12 non-endorsed candidates, the result was a universal loss. The US House experienced a slightly more optimistic win/loss split, with a total of 49 endorsed candidates having 30 wins and 19 losses. However, in no way do the pessimistic results of non-endorsed candidates imply the decisiveness of a Trump endorsement; Trump has a high propensity to endorse candidates which are likely to succeed in the general election regardless of his endorsement due to a variety of fundamental factors, the most important of which include PVI (political leaning of state) and incumbency. It thus speaks volumes that none of the non-Trump endorsed candidates for Senate were incumbents before the 2018 midterm Senate election (historically, challengers have struggled to displace incumbents), representing an average PVI of -9, more than 15 net percentage points than the average PVI of endorsed candidates' states and 12 less than the national average. Results from our support vector machine model's weight vector serve to justify this.

Variable	Weight	Percentage Contribution
Trump Endorsed?	0.419	11.0%
Incumbent?	0.518	13.6%
PVI	1.44	37.8%
Percentage White	-0.00772	0.203%
Percentage w/ Bachelor's Degree	-0.956	25.1%
Percentage Aged 65 or Over	-0.192	5.04%
Median Income	0.278	7.30%

We evaluated the accuracy of our model with a confusion matrix as follows, obtaining an accuracy of 86%:

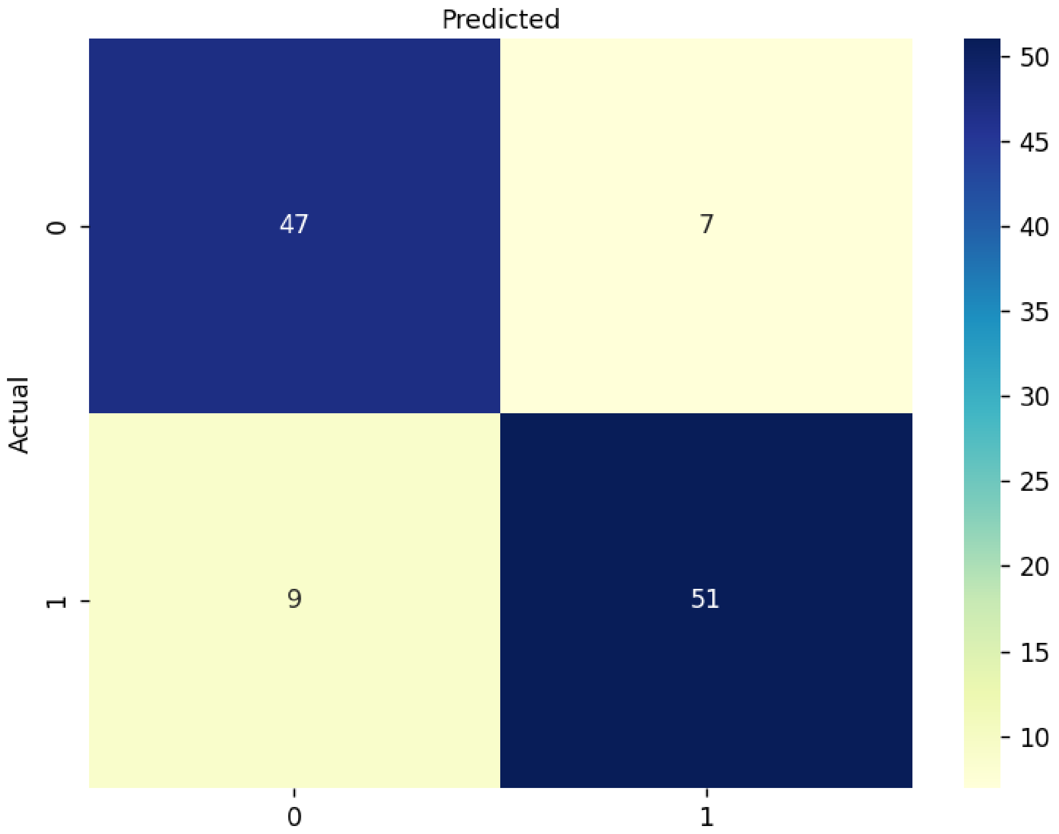


Fig. 15: Confusion Matrix for Trump Endorsement SVM

To assess the model further, we applied data from only the 2020 midterm election into the model and compared the predicted results to the actual outcomes:

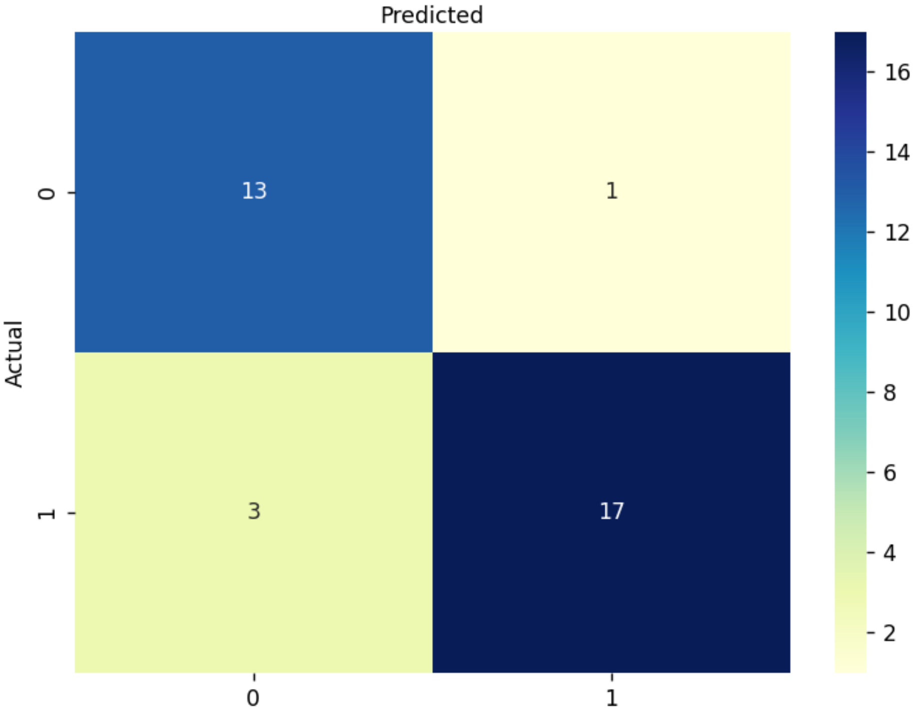


Fig. 16: Confusion Matrix for 2020 Trump Endorsement SVM

A total of 34 senate elections occurred in the 2020 midterms, with 30 correct predictions and 4 incorrect predictions by our model trained on 2018 data for an 88% accuracy. Trump's endorsements, though non-trivial, play a small role in determining the outcome of the elections (11% contribution), while PVI, unsurprisingly, is the best predictor and contributor (37.8%), closely followed by educational attainment (25.1%), which has an inverse relationship with Trump support. Trump endorsements are also outweighed by incumbency, which contributes 13.6% of the election outcome and also serves as an effective predictor of the result.

3.3 How Powerful Are Sanders Endorsements?

A total of 45 candidates in the House and Senate were endorsed by Bernie Sanders in the 2018 and 2020 midterms combined, including 8 Senators and 37 Congresspeople; 31 Senators during the same period did not receive Sanders endorsements. Out of all Sanders-endorsed candidates advancing to general elections, 27 won their elections while 18 lost, a 60% success rate; in contrast, non-Sanders endorsed Senators experienced 21 wins and 10 losses, a 68% success rate. While Trump's endorsements exhibit a tendency towards "safe" candidates in firmly Republican states, Sanders' endorsements occur regardless of political inclination or environment in districts and are linked to individual political ideologies, potentially leading to Sanders-supported progressives in deeply Republican districts losing their elections. Applying the SVM model, we find the following weights for each variable in contributing towards electoral outcomes:

Variable	Weight	Percentage Contribution
Sanders Endorsed?	0.261	6.17%
Incumbent?	0.863	20.4%
PVI	-1.15	27.2%
Percentage White	-0.333	7.87%
Percentage w/ Bachelor's Degree	0.562	13.3%
Percentage Aged 65 or Over	0.389	9.20%
Median Income	-0.672	15.9%

with the following accuracies:

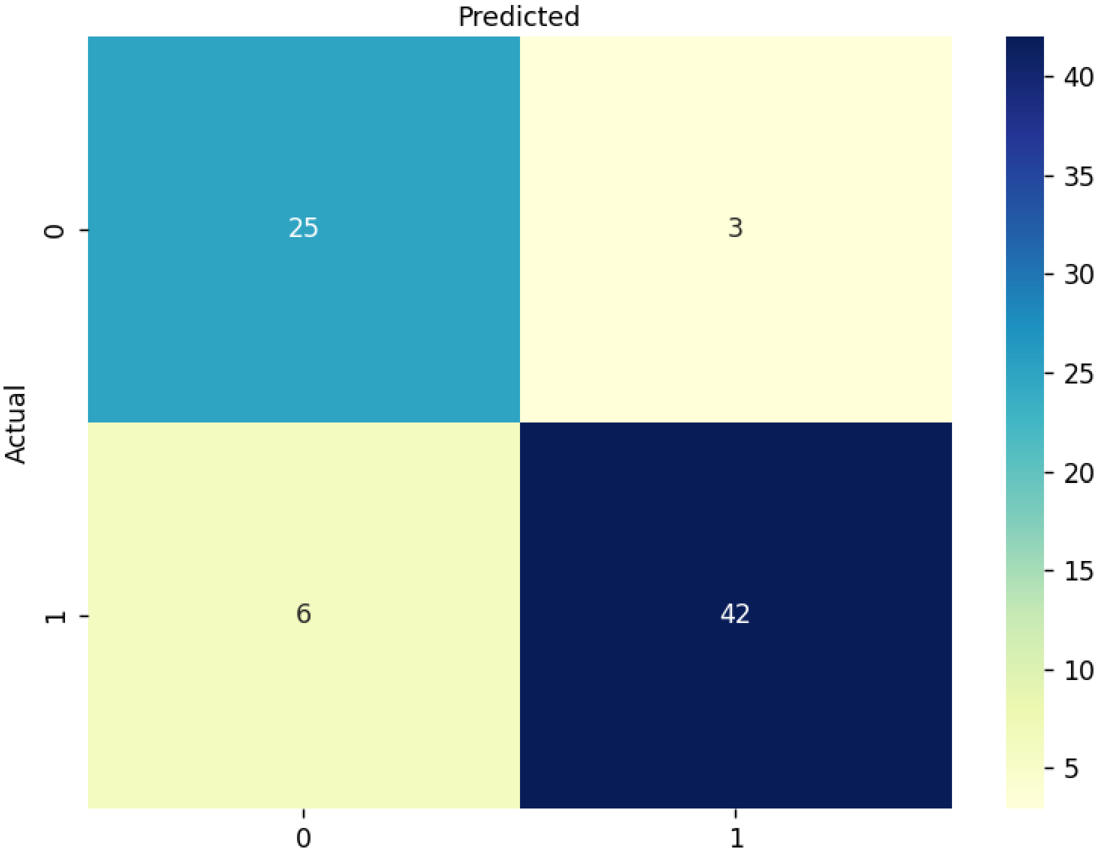


Fig. 17: Confusion Matrix for Sanders Endorsements SVM

We thus conclude that in both instances of Trump and Sanders' endorsements, the endorsements themselves, though non-negligible and positive, contribute an insignificant amount towards the electoral outcome of the candidate. Far better predictors are the PVI, which accounts for approximately 37% of the outcome in both cases, and socioeconomic factors such as educational attainment and median income, with both factors combined accounting for more than 30% of the outcome. The incumbency of the candidate also plays a major role (10 to 20%), in accordance with a growing trend of Congressional stagnation where well over 90% of incumbents are reelected (Murse, 2020).

3.4 Predicting the 2022 Midterms Through Trump Endorsements

Political polarization in the US has reached unprecedented levels during the past few years, with the nation entrenched deeper within fundamental political divisions, not only between Republicans and Democrats borne of an inability to accept democratically produced election results, but between Trump-endorsed and Trump-denounced candidates. In the 2022 election, Trump has once again embraced his role as the most prominent figure in the Republican party, endorsing a total of 22 candidates in the Senate alone; across both chambers of Congress, 40 candidates have echoed Trump's claims of election fraud and have firmly aligned themselves with the former president. The 2022 midterms will act as the greatest test to Trump's continued dominance over the Republican party; using aggregated data from 2018 and 2020, we will attempt to predict the outcome of the 2022 Senate elections with a particular emphasis on Trump endorsements.

3.4.1 The 2022 Prediction Model

Training our previous logistic regression and support vector machine model on data obtained from both the 2018 and 2020 midterm election cycles, regression coefficients for endorsement likelihood gives:

Variable	Coefficient/Weight	Percentage Weight	Weight in 2018
PVI	0.302	37.5%	55.0%
Percentage White	-0.0434	5.39%	0.560%
Incumbent?	0.0102	1.24%	1.31%
Percentage w/ Bachelor's Degree	-0.191	23.7%	28.2%
Percentage Aged 65 or Over	-0.258	32.1%	14.7%
Median Income	0.000151	0.0188%	0.0237%

As shown, the PVI of a state remains the most decisive factor in determining whether or not a state's candidate will receive a Trump endorsement; so do demographic factors like age distribution and education level which significantly affect Trump support within the state. Evaluating whether or not the same patterns hold within 2022, we find that our model retains an 85% accuracy when predicting 2022 Senate endorsements by Trump compared to actual results:

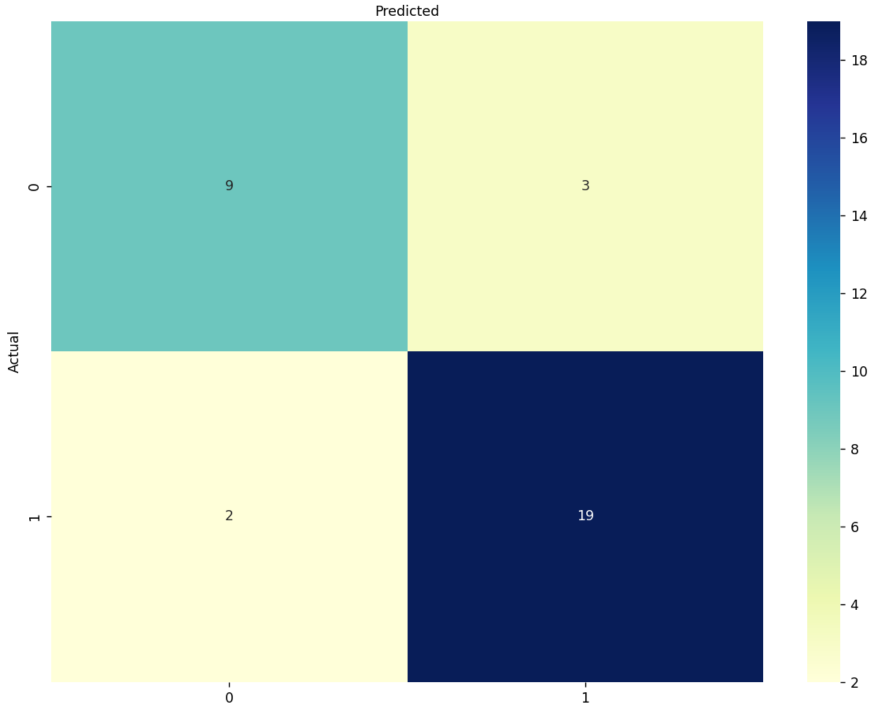


Fig. 18: Confusion Matrix for 2022 Trump Endorsement Predictions

This implies that the trends existing in previous years of Trump's inclination towards "safe" candidates in highly Republican states with higher levels of Trump support still holds in 2022, albeit to a slightly diminished extent as Trump endorses more frequently in swing states with less favorable PVIs. Our electoral outcome SVM model is identical to the one described above, with a 36.8% weight assigned to PVI, a 25.1% weight assigned to educational attainment, and a non-trivial but relatively insignificant 11% weight assigned to Trump endorsements:

Variable	Weight	Percentage Contribution
Trump Endorsed?	0.419	11.0%
Incumbent?	0.518	13.6%
PVI	1.44	37.8%
Percentage White	-0.0077 2	0.203%
Percentage w/ Bachelor's Degree	-0.956	25.1%
Percentage Aged 65 or Over	-0.192	5.04%
Median Income	0.278	7.30%

Due to the entirely binary nature of the classification model combined with the fact that no source of actual election results exists for cross-validation, we introduce Platt scaling on our SVM results as a calibrator to assign probabilities to our predicted outcomes (Lin, 2007). Platt scaling was introduced chiefly to convert SVM classifications into probabilistic outcomes, and operate off of the decision function $f(x)$ of SVMs (Platt, 1999):

$$f(x) = w \cdot x + b$$

where w is the weight vector, x is the input vector and b is the bias constant. This decision function represents the distance between the optimal hyperplane and the input vector; thus, as further distance represents greater certainty for one classification over the other, Platt scaling fits the decision function against the classification of the point based on a logistic regression (sigmoid curve):

$$P(x = 1) = \frac{1}{1+e^{-Af(x)+B}}$$

generating probabilities for x to be classified as 1 (victory) and 0 (defeat). This probabilistic calibration allows us to identify potential swing states and secure states, potentially pointing towards the uncertainties existing in our results. Our 2022 midterm Senate election predictions, alongside our victory probabilities generated by Platt fitting compared to FiveThirtyEight's predicted probabilities, are as follows:

Candidate	State	Predicted Outcome	Predicted Probability	538
Britt	Alabama	Win	98.5%	99%
Tshibaka	Alaska	Win	77.4%	68%
Masters	Arizona	Win	74.3%	26%
Boozman	Arkansas	Win	99.2%	99%
Meuser	California	Lose	5.00%	1%
O'Dea	Colorado	Lose	12.1%	10%
Levy	Connecticut	Lose	4.50%	1%
Rubio	Florida	Win	76.1%	86%
Walker	Georgia	Win	73.7%	51%
McDermott	Hawaii	Lose	2.59%	1%
Crapo	Idaho	Win	99.0%	99%
Salvi	Illinois	Lose	15.1%	1%
Young	Indiana	Win	95.6%	99%
Grassley	Iowa	Win	82.3%	97%
Moran	Kansas	Win	88.8%	99%
Paul	Kentucky	Win	99.1%	99%
Chaffee	Maryland	Lose	1.00%	1%
Schmitt	Missouri	Win	92.8%	98%
Laxalt	Nevada	Win	77.2%	37%
Pinion	New York	Lose	6.86%	1%

Budd	North Carolina	Win	76.6%	63%
Hoeven	North Dakota	Win	98.1%	99%
Vance	Ohio	Win	86.7%	72%
Mullin	Oklahoma	Win	99.2%	99%
Lankford	Oklahoma	Win	99.2%	99%
Perkins	Oregon	Lose	17.6%	1%
Oz	Pennsylvania	Lose	24.0%	20%
Scott	South Carolina	Win	92.1%	99%
Thune	South Dakota	Win	97.1%	99%
Lee	Utah	Win	89.8%	94%
Malloy	Vermont	Lose	3.08%	2%
Smiley	Washington	Lose	10.7%	3%
Johnson	Wisconsin	Win	68.7%	51%

The model predicts similar results to expert predictions and other predictive models such as FiveThirtyEight's models; some deviation occurs, however, in states where the model expresses less certainty, there is some deviation in the predicted probabilities and results. Using 75% as a margin, the model forecasts Arizona, Georgia, and Wisconsin as potential swing states; it predicts similar results to FiveThirtyEight for Georgia and Wisconsin (though with higher certainty), but deviates significantly for Arizona and fails to identify Nevada as a swing state. The following is a map, color-coded from blue (0% victory confidence for Republican candidate) to red (100% confidence), representing the probabilistic predictions of our model for the 2022 midterm Senate elections.

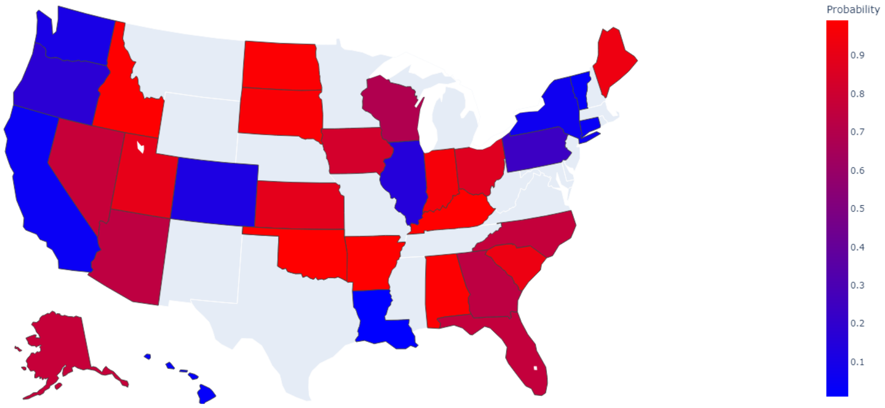


Fig. 19. Victory Probability Chloropleth Map for 2022 Senate Midterm Predictions

3.4.2 Simulating the 2022 Midterm Elections

Using the probabilities generated from our model outlined above, we simulated the election a total of 1000 times to predict the range of potential outcomes in terms of seats won by each party as well as their corresponding likelihoods. As the Senate currently stands, the two parties are deadlocked in an even 50-50 split (with Republicans holding 50 seats and Democrats holding 48 seats alongside two independent senators caucusing with the Democrats); due to the deciding vote of the Vice President under tiebreak scenarios, however, the Democrats will control the Senate if a further 50-50 balance occurs after the 2022 midterms. A total of 35 Senate seats are in contention in the 2022 midterms (21 Republican and 14 Democratic/Democrat-aligned), with a total of 22 Republican victories thus required to attain a majority in the Senate. Our results, alongside their corresponding probabilities are plotted in the following frequency bar chart:

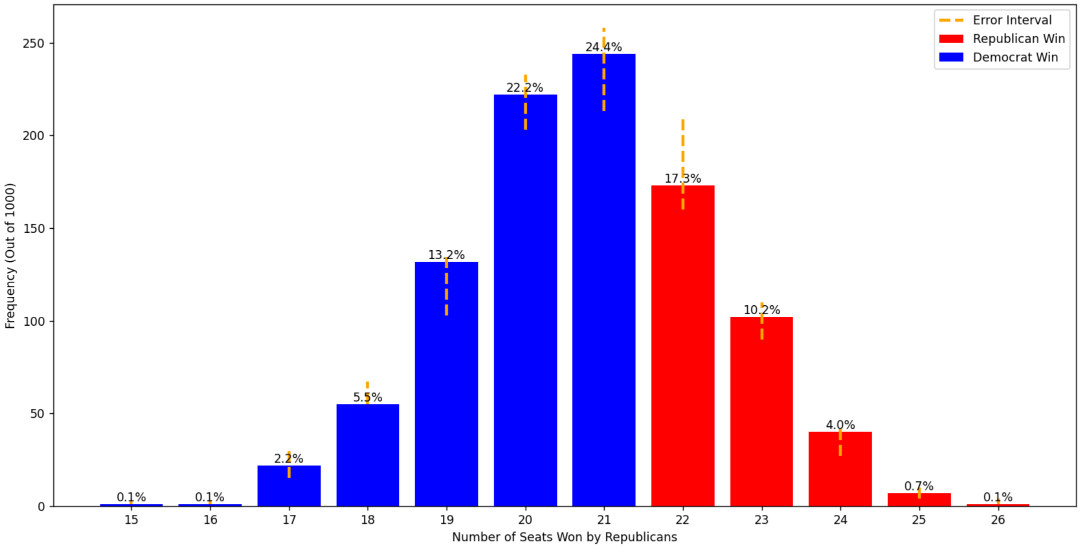


Fig. 20. Potential Outcomes and Corresponding Probabilities for 2022 Senate Midterms

The potential outcomes range from 15 to 26 Republican seats in the Senate, out of a total of 35; in a total of 77.1% of scenarios, the Republicans will win between 19 and 22 seats. Conclusively, the aggregate probability for Democrats to retain control over the Senate is 67.7%, while the Republicans control the senate in 32.3% of all scenarios.

4. Limitations and Future Research

This paper presents a logistic regression framework for predicting whether or not a candidate in the US midterm elections will receive endorsements by major political figures such as Bernie Sanders and Donald Trump, and illustrates the dichotomy of ideological and pragmatic endorsements through comparing the accuracy of a fundamental model and a progressivity score model on Sanders and Trump endorsements. However, this framework has several limitations.

First and foremost, general elections are a poor descriptor of the true impacts of endorsements. Though endorsements may not contribute much to general elections - a trend validated by previous polling and research, and supported by the phenomenon that voters from the opposing party may also be galvanized to vote by the endorsement (e.g. Trump's) - they can play a major role in primary elections, in which Trump's supported candidates have won 97% of the time. This is due to only voters affiliated with a certain party participating in primaries, making Trump or Sanders or the influence of any political figure far greater. Future research could attempt to apply the same binary classification model alongside fundamental factors in US midterm primary elections.

Furthermore, the fundamental variables selected in the model - age, race, and socioeconomic factors - may not be the most optimal combination of factors. In fact, our results have yielded some evidence that may suggest how some of these variables will negatively affect the accuracy of the predictions made. When all fundamental variables were excluded in favor of progressivity in predicting Bernie endorsements, accuracy jumped from 67% to 80%; when age was excluded and all other fundamental factors were included, accuracy also increased to 70%. Age thus appears to be less relevant than other factors to the predictions, and may have decreased the overall accuracy of the model. As the model combines data from Congressional districts and states, several Congressional districts with small populations also serve as major outliers with high incomes, educational attainment, or extreme PVIs (e.g. California's 17th district has a median income of $142,408, just shy of triple the national average, and a Bachelor's attainment rate of 60%); future research could benefit from using a weighting system considering the population of each district to prevent massive outliers from affecting accuracy.

In terms of accuracy, the model's described trends and predictions do achieve fairly satisfying results (>80%, with Trump's endorsement model close to 90% in both 2018-2020 and 2022), but crucially lacks reliable predictions in swing states such as Nevada or Georgia. Though it is currently impossible to determine the validity of the 2022 predictions generated by the model, all inaccuracies of the model in 2020 and 2018 stemmed from states identified as swing states. The 2022 prediction model also did not consider Sanders endorsements due to a lack of endorsed candidates. Additionally, the probabilities generated by the Platt scaling component of the model are occasionally too moderate; For instance, Republican candidate for US Senate in Illinois, Salvi, is unlikely to have a 15.1% probability of victory and is nearly certain to lose the election. This gives rise to potential scenarios in the election simulation where firmly Democratic or Republican states flip their seats. Future models can improve upon the accuracy of our results by considering a range of individualized variables for each candidate (e.g. recent scandals, events, or donations), as fundamental variables are static over time and do not reflect important political trends. Furthermore, only Trump and Sanders were considered in this paper as two case studies of diametrically opposed endorsement patterns; in future research, the same model may be applied to study and examine patterns in other political figures' endorsements, such as Obama or Biden, and classify them as pragmatic or ideological.

5. Conclusion

Through analyzing the endorsement patterns of Donald Trump and Bernie Sanders in the 2018 and 2020 midterms, this paper broadly classifies endorsements in the US by major political figures as either pragmatic (well-predicted by fundamental variables such as age, race, income, educational attainment and especially political inclination of district) or ideological (well-predicted by the individual candidate's political ideology). Using a logistic regression model, we found that Trump's endorsements are primarily driven by the political leaning and demographic factors of the state rather than individual candidates, while Sanders' endorsements were better predicted by the progressivity of individual candidates. Applying a support vector machine model, we sought to determine the contribution of each fundamental variable as well as the endorsements of Trump and Sanders towards electoral outcomes in US midterms in 2018 and 2020, and found that while endorsements provide a positive and non-negligible boost towards electoral results (11% and 7% respectively for Trump and Sanders), PVI (political leaning) and socioeconomic variables (average income and educational attainment) combined contribute over 75% of the result, with incumbency of the candidate contributing the majority of the remaining 25%. Finally, we predicted outcomes of the 2022 Senate elections using our SVM model trained on 2018 and 2020 data, considering fundamental variables alongside Trump endorsements; generating probabilities for Republican victories in each state and simulating the Senate election 1000 times, we found a 67.7% aggregate probability for Democrats to retain the senate, with a range of 15 to 26 Republican victories possible.

References

Ballard, Andrew O., et al. “Be Careful What You Wish for: The Impacts of President Trump’s Midterm Endorsements.” *Legislative Studies Quarterly*, vol. 46, no. 2, 2020, pp. 459–491., https://doi.org/10.1111/lsq.12284.

Ballotpedia Editor. “Endorsements by Bernie Sanders.” *Ballotpedia*, 2 Sept. 2022, https://ballotpedia.org/Endorsements_by_Bernie_Sanders.

Ballotpedia Editor. “Endorsements by Donald Trump." *Ballotpedia*, 28 Sept. 2022, https://ballotpedia.org/Endorsements_by_Donald_Trump.

Cook Political Report. “Introducing the 2017 Cook Political Report Partisan Voter Index.” *Cook Political Report*, 2017, https://www.cookpolitical.com/introducing-2017-cook-political-report-partisan-voter-index.

Cortes, Corinna, and Vladimir Vapnik. “Support-Vector Networks.” *Machine Learning*, vol. 20, no. 3, 1995, pp. 273–297., https://doi.org/10.1007/bf00994018.

Lin, Hsuan-Tien, et al. “A Note on Platt’s Probabilistic Outputs for Support Vector Machines.” *Machine Learning*, vol. 68, no. 3, 2007, pp. 267–276., https://doi.org/10.1007/s10994-007-5018-6.

Platt, John. “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods.” 26 Mar. 1999.

Ester, Martin, et al. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.” *Association for the Advancement of Artificial Intelligence*, 1996, pp. 226–231. 1., https://doi.org/10.5120/739-1038.

Sharma, Ankush, and Amit Sharma. “KNN-DBSCAN: Using K-Nearest Neighbor Information for

Parameter-Free Density Based Clustering." *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, 2017, https://doi.org/10.1109/icicict1.2017.8342664.

US Census Bureau. "Profile." *Census Reporter: Making Census Data Easy to Use*, https://censusreporter.org/.

US Census Bureau. "Voting Statistics: State Electorate Profiles." *Census.gov*, 8 Oct. 2021, https://www.census.gov/library/visualizations/2016/comm/electorate-profiles.html.

Murse, Tom. "Do Members of Congress Ever Lose Re-Election?" ThoughtCo, *ThoughtCo*, 10 Dec. 2020, https://www.thoughtco.com/do-congressmen-ever-lose-re-election-3367511.