Biden or Trump?

Predicting the Outcome of Votes Based on Donations Data in Battleground States

Abstract

Voters and lawmakers in the United States are consistently engaged in debate regarding the influence of political campaign contributions. This analysis sheds light on the predictive power of political donations on presidential election outcomes in key battleground states in the United States and the insights they provide. We contribute quantitative arguments to this discussion by using machine learning to understand how donations to political campaigns inform voting in counties and states, based on datasets collected from publicly available resources. As active citizens, our group is invested in evaluating how events similar in magnitude have created shifts in political affiliation.

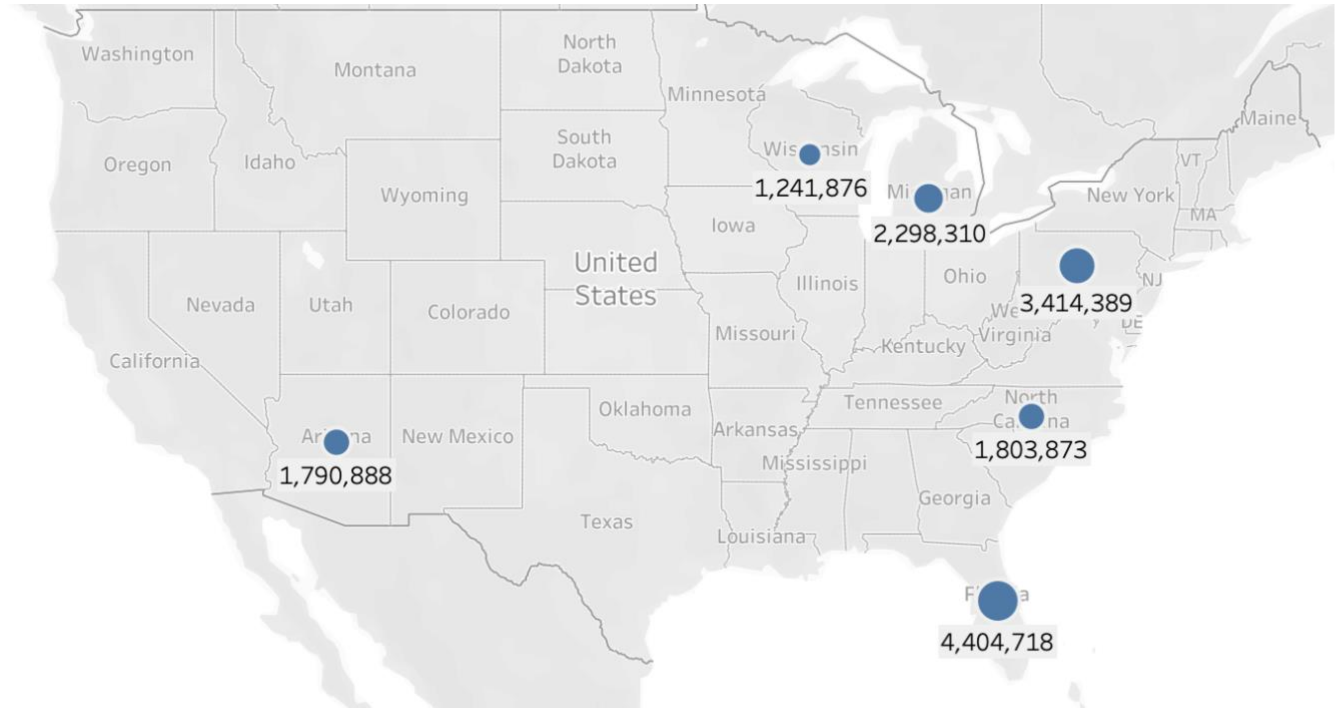
We used both a logistic regression model paired with an unsupervised clustering model in order to predict who will win the election in six swing states: Arizona (AZ), Florida (FL), Michigan (MI), North Carolina (NC), Pennsylvania (PA), and Wisconsin (WI). We chose to use these models, given that logistic regression analysis and unsupervised machine learning can be used to predict the outcome of binary dependent variables. We also perform a linear regression analysis on aggregated number of votes based on donation, along with donations based on other non-binary features such as city, state, zip code, and employer.

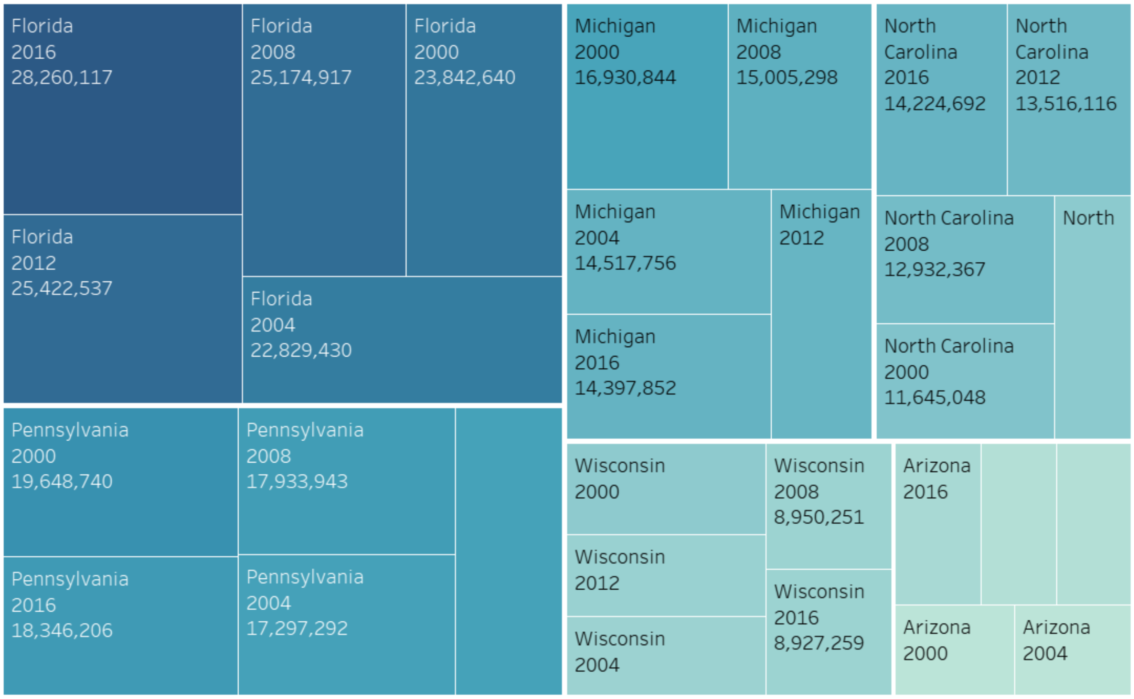
The main contribution of this study is a careful analysis of the predictive power of donations in battleground states in determining presidential election outcomes. First, we provide an overview of the questions guiding our analysis. We describe the analytical processes involved in our analysis, including data requirements specification, collection, processing, and cleaning. We then provide a discussion on our machine learning models, and end with the results and takeaways from our analysis.

Introduction

The billions of dollars raised and spent during federal elections indicate why political campaign donations play a crucial role in the fate of a political candidate. Campaign donations help political candidates fund events, proliferate advertising, and assess the impact of their campaign in various states. Many studies explore the extent of the influence of money on the political process, and in this study, we aim to determine the extent of the influence of donations made by residents of battleground states on electional outcomes of those states: whether more donations in battleground states make them swing red or blue.

We focus our search on battleground states in the United States because of the impact they have in federal elections. Battleground states are those states in which neither major political party holds a lock on the outcome of presidential elections, and whose electoral votes have a high probability of being the deciding, or ‘swing' factor in a presidential election. In our analysis, we use the 2020 considerations for battleground states: Arizona (AZ), Florida (FL), Michigan (MI), North Carolina (NC), Pennsylvania (PA), and Wisconsin (WI). Given that these states shift back and forth between candidates of both political parties and makeup nearly 20% of all donors, electoral college votes, and delegates year on year, candidates focus events and fundraising campaigns in these states. We replicate our analysis across years between 2000 and 2020, mapping political campaign donations and federal election outcomes to build a predictive clustering model that helps us understand which party may win these key states in 2020.





Our model captures the predictive power of political donations on presidential election outcomes. In order to predict how political affiliations in key battleground states may change and inform voting as a result, some problems statements we address include:

> Is there a significant transfer of donations from the Republican party to Democratic during times of crisis or vice versa?
>
> How do demographics (age, ethnicity, geographical delineations, employment status, employment, education, and party affiliation) in each county play into how constituents engage with donations?
>
> Do these demographics have predictive power in presidential elections?
>
> Do donations have predictive power in presidential elections?

Through our analysis, we find that donations to political parties in swing states can be predictive of election outcomes in those states. The following sections outline our analytic approach to building our predictive model.

Data Analysis

To build our model, we first began by performing the standard steps of the data analysis process. This included steps to clarify data preprocessing and collection, cleaning, feature engineering, and finally, our data analysis.

Data Preprocessing and Collection

In order to understand the variety of data we would require to construct our model, we consulted our guiding questions. As many of the inputs in our model would require demographic and donations data, we consulted publicly available databases and resources that provided raw voting data, Census Bureau data for demographic and unemployment features, and FEC donations and campaign financing datasets. For our analysis, we use the following datasets:

	Features	Sources
Voting Data	Election outcomes during various crisis years (Dotcom bubble burst, financial crisis, COVID19), presidential candidate and party data	Electoral College Votes, Votes by State
Demographics	Urban/Suburban split, ethnicity, county information, zip codes of counties	Health Metrics, Census Bureau Data
Unemployment Data	Year, month, season	Census Bureau Data
Donations and Contributions Data	List of individual and corporate contributions at a county-level made to candidates, in battleground states	FEC Campaign Finance Data, FEC General Data

Data Cleaning

Upon extracting the data from the publicly available sources, we began processing and cleaning our datasets.

Logistic Regression Model: After gathering data from master files in publicly available databases, we proceeded to convert data used for our logistic regression models into appropriate formatting using le encoder. We proceeded to create various tables and processed the data in Jupyter Notebook and Google Cloud. This gave us a better understanding of each data element, along with what information we need to group, focus on and needed to be removed. After parsing the data, we optimized it for our project. We concluded that looking at previous voter information, coupled with donation details and demographics would provide insights on our focus question: donations affecting presidential elections in swing states.

Unsupervised Clustering: Using values gathered from Census Bureau databases, all data was loaded into a Postgres Database. It was then extracted into a data frame using Jupyter Notebook. For null values, we performed analyses and concluded that 0 would be an appropriate filler value, resulting in parsing null data to complete data with 0 values. Once in the model, data was scaled to avoid skewing results.

All cleaned data was aggregated in several tables, as displayed in the table below:

	Description	Table Name	Rows	Columns
1.	Donors on selected states	agg_county_donors	6,748	12
2.	Voters on selected states	agg_county_votes	2,020	14
3.	Births and deaths	birth_death_rate	305	6

4. Presidential candidates since 2000	candidates	141,375	17
5. Committee presidential donations	committees	141,375	15
6. Donations for presidential votes	donations	14,954,054	21
7. Educational levels by state	education	3,283	47
8. Health information	health_metrics	3,193	507
9. County postal codes	postal_codes	52,889	5
10. Presidential votes on selected states	pres_votes_6T	6,464	11
11. Donations for presidential votes	six_state_donations	14,954,054	21
12. Unemployment rates	unemployment	404	114

Feature Engineering and Selection

Carefully selected features in a machine learning model ensure flexible, simple models with accurate results. Because features in data directly influence the predictive model and its results, we spent significant time exploring our cleaned datasets using Jupyter Notebook and Google Colab. This process involved building a better understanding of each data element, along with establishing data that needed to be grouped in order to create effective features. We concluded that values of found in donations datasets, paired with demographics in counties where donations were observed would provide insight on how these donations may have affected presidential elections in those states in prior years, and whether those behaviors could be predicted in 2020. We performed the following joins on datasets to create new data frames for analysis:

County: Links were performed to other datasets in order to run an aggregated data frame. These included unemployment, urban population density, and rural population density datasets.

Donations: 4-year segments were created to represent each federal election cycle, and summations were performed in order to move from a city-level to county-level view. These counties were then mapped within the 6 concerned states. Additionally, committee IDs were identified in order to map to whether the committees were Democratic or Republican. The joins on these separate datasets helped create one cohesive data frame that filtered county-wide donations into 'Red' or 'Blue' votes.

Voter: Individual donations datasets (agg_county donors) were keyed to voter database (agg_county_votes), accounting for and displaying state and election years associated with each.

ETL: Features were further engineered for simplicity and accuracy by removing approximately 10 irrelevant parameters

Logistic and Linear Regression Models: Key features in this model come from the donor table (agg_county_donors) for the six swing states. We selected the committee city, state, zip code (postal_codes), transaction amounts (donations), along with donor's employer and donor's occupation.

***Unsupervised Clustering**:* The principal component analysis (PCA) shows that only 3 features would ultimately drive segmentation. From there, an analysis of model inertia showed a selection between 3 and 5 clusters. We processed data in various tables using Jupyter Notebook and Google Cloud.

Model Methodology and Summary

Our predictive model performs a linear and logistic regression analysis, along with unsupervised clustering and a classification model. Understanding our decision to proceed with these models involves first discussing the considerations and limitations involved with each model.

Logistic Regression*:* As logistic regression analysis is used to predict a binary dependent variable, we use logistic regression analysis to act as predictive of party classifiers such as the win of a Democrat or Republican in a battleground state. Logistic regression models work well for this portion of our analysis, because the trained weights provide inferences about the importance of each feature. Logistic regression can be used to find the relationship between the features, and updates to the model can be performed easily unlike support-vector machine learning or decision trees. The outputs of logistic regression models are also well calibrated to probabilities. It is less prone to overfitting in a low dimensional dataset, and therefore provides the perfect model for us to act as predictive of 'red' or 'blue' wins in each state. We acknowledge that there are limitations to this model: on high dimensional datasets that are sensitive to outliers, logistic regression models may lead to overfitting due to the model's tendency to predict probabilistic outcomes based on independent features. However, given that party classifier data is not highly dimensional, a logistic regression model suits our predictive model for party classifiers.

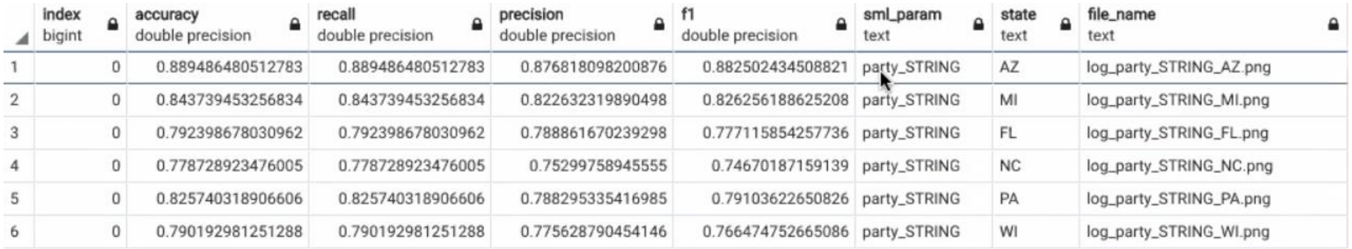
Linear Regression*:* As this linear regression models serve as a predictive method for continuous variables, we performed a linear regression in 2 parts: 1) on aggregate number of votes based on donations and; 2) on donations based on other non-binary features such as city, state, zip code and employer. We chose a linear regression model due to its ease of implementation and interpretation. Because our prior data analysis allowed us to understand our feature set closely, our linear regression model allowed us to accurately represent the relationship between the various independent and dependent variables among our datasets. While linear regression models are susceptible to overfitting, we implemented regularization techniques to reduce this impact.

Unsupervised Clustering*:* After performing a neural network on our existing dataset, we found that our results indicated a 0.98 fit in the deep learning model. This led us to believe that a neural network model was likely overfitting our data. Based on this result, our data, and out predictive problem statement, we believe we may be better suited to using unsupervised clustering models such as k-means clustering. Easy to implement, capable to scaling to our large dataset, and with its ability to generalize to various clusters, we believe that unsupervised clustering suits our machine learning model best. The goal of the clustering model is to segment individual counties based upon a range of unemployment and health metrics. This model could be extended further to help presidential candidates assess the demographics of each county and state, how donations are impacted as a result, and how to tailor their efforts in the county or state to best serve their purpose of election.

In order to ensure that we can map the county-level insights gathered from our models to overall, 'winner-takes-all' results at the state level, we also introduce a classification model. This model predicts the 'Democrat' or 'Republican' result using the inputs of county-level results in each state.

For our logistic and linear machine learning models, we split testing and training sets based on what is considered 'industry standard' in data visualization: our training and testing data is split in 75% training, and 25% testing.

We measured the success of all of our models based on the success of our key predictive model: the logistic regression model. Since the logistic regression model predicts the 'Democrat or Republican' outcome in each state, we look at precision and recall scores collected in order to measure the success of the various models that feed into it. Based on this analysis of precision and recall scores, we believe this model will help candidates target donors by accurately predicting which party they will donate to, and how those donations will be predictive of the outcomes of elections in the state.



	index bigint	accuracy double precision	recall double precision	precision double precision	f1 double precision	sml_param text	state text	file_name text
1	0	0.889486480512783	0.889486480512783	0.876818098200876	0.882502434508821	party_STRING	AZ	log_party_STRING_AZ.png
2	0	0.843739453256834	0.843739453256834	0.822632319890498	0.826256188625208	party_STRING	MI	log_party_STRING_MI.png
3	0	0.792398678030962	0.792398678030962	0.788861670239298	0.777115854257736	party_STRING	FL	log_party_STRING_FL.png
4	0	0.778728923476005	0.778728923476005	0.75299758945555	0.74670187159139	party_STRING	NC	log_party_STRING_NC.png
5	0	0.825740318906606	0.825740318906606	0.788295335416985	0.79103622650826	party_STRING	PA	log_party_STRING_PA.png
6	0	0.790192981251288	0.790192981251288	0.775628790454146	0.766474752665086	party_STRING	WI	log_party_STRING_WI.png

Results

Based on our observations of the model, its structure, and its predictive power, we believe it could be refined and used to:

- **Predict potential Congressional votes**: Our predictive modeling could be used to provide insights and informative studies on how elected officials in battleground states may vote on Congressional bills. <Insert some other studies that have been performed on this subject, provide references>
- **Inform campaign financing efforts and impact**: Our model may help those working in political campaigns and finance offices to tailor their efforts. This use case could also lend itself to providing a source in studying the impact of campaign financing on presidential outcomes.
- **How voting is affected during crisis events**: Our model may help political experts understand and account for how voting activity may be impacted during times of economic or social crises.