

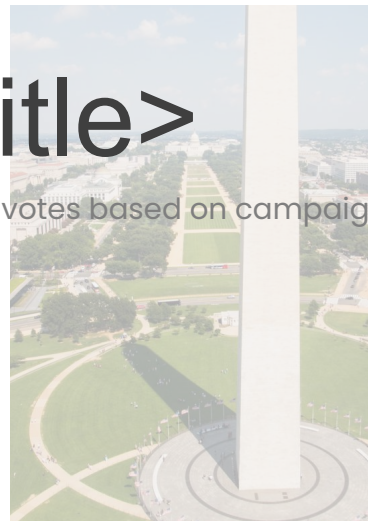


<Title>

Predicting election votes based on campaign donations

Paul Bryzek, Doris Cohen,

Josh Kedzierski, Kanika Singh



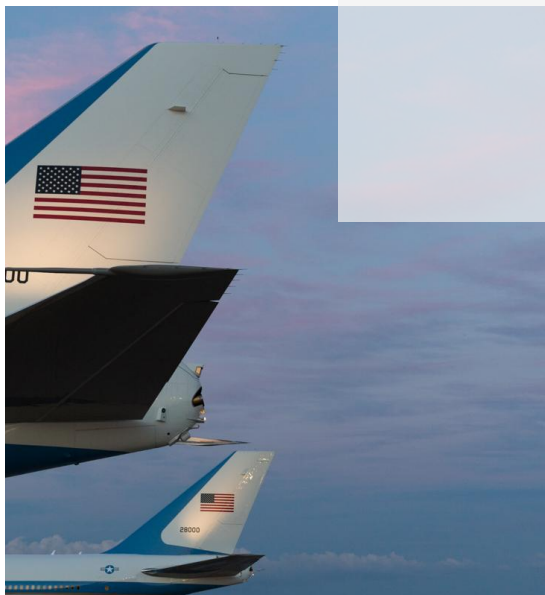


Project Summary

This project analyzes the predictive power of political donations on presidential election outcomes in key battleground states in the United States. We have built several predictive models using both supervised and unsupervised machine learning to understand how donations to political campaigns inform voting in counties and states. We will also look at times of economic and social crises and how it may impact and shift the parties to which voters choose to donate.

[Dashboard](#)

[GitHub](#)





Michigan

lorem ipsum dolor sit
consectetuer



North Carolina

lorem ipsum dolor sit
consectetuer



Arizona

lorem ipsum dolor sit
consectetuer



Florida

lorem ipsum dolor sit
consectetuer



Pennsylvania

lorem ipsum dolor sit
consectetuer



Wisconsin

lorem ipsum dolor sit
consectetuer



Our study

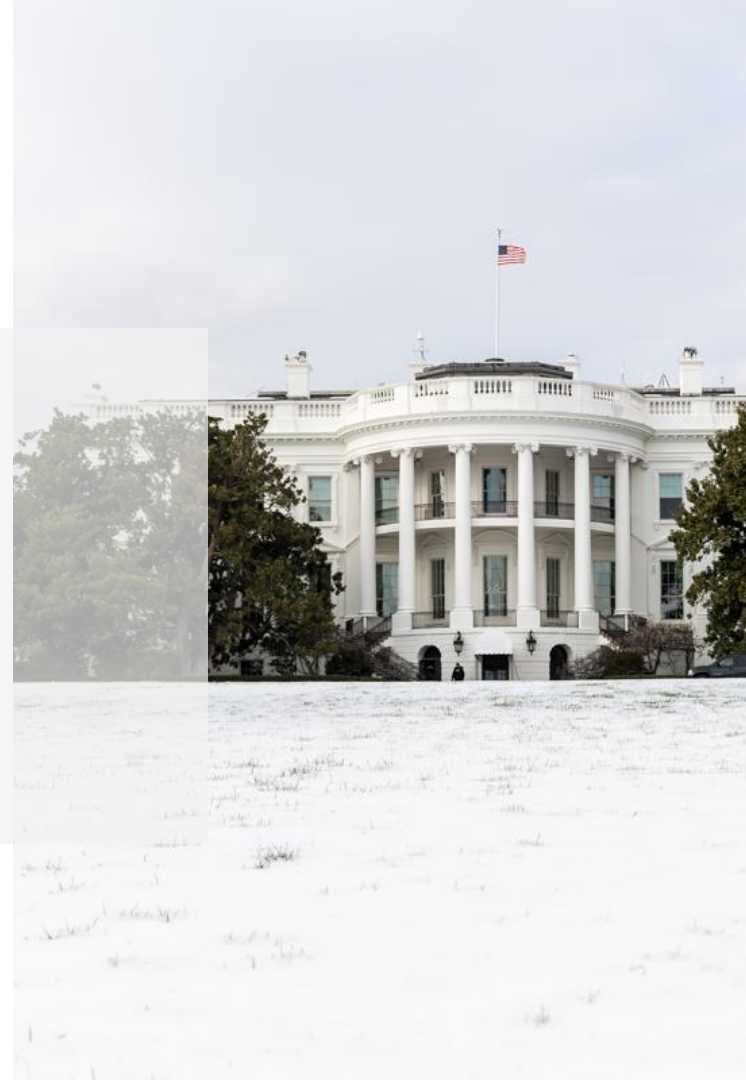
Our model captures how crises between 2000 & 2020 may have impacted campaign donations in order to predict how political affiliations in key battleground states may change and inform voting as a result. Some problems statements we address include:

- 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?

A decorative graphic consisting of four plus signs of varying sizes and colors (red and blue) arranged in a cluster to the left of the title.

Why is this important?

We found that this topic could have impactful predictive power in understanding how times of crises can influence the outcomes of presidential elections. The state of the pandemic in the US paired with civil unrest throughout the country pose a dynamic background for the upcoming presidential elections. As active citizens, our group is invested in evaluating how events similar in magnitude have created shifts in political affiliation.





Practical applications

Some ways in which we believe our predictive modeling would prove beneficial include:



Predicting Congressional Votes

Our predictive modeling to give insight into how elected officials in battleground states may vote on congressional bills.



Campaign Financing

Our model may help those working in political campaigns and finance tailor their efforts.



Voting in Crises

Our model may help political experts understand and account for how voting activity may be impacted during times of economic or social crisis.



Dataset Overview

For our analysis, we used publicly available datasets
that evaluate the following features:

Voting Data



Crisis: Dotcom
Bubble, 9/11



Crisis: Financial
Crisis



Crisis: COVID-19

[Electoral College Votes](#),
[Votes by State](#)

Demographics



Urban vs Rural



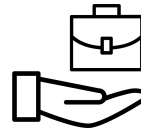
Ethnicity



County

[Health Metrics](#), [Census
Bureau Data](#)

Unemployment Data



Year



Month



Season

[Census Bureau Data](#)

Donations



6 States



County-specific



Bipartisan

[FEC Campaign Finance
Data](#), [FEC General Data](#)



Feature Engineering

Upon processing our collected data and optimizing it for our project, we concluded that mapping past voter information on donations data and demographics would provide insights on how donations affect presidential elections in swing states. We conducted the following joins to create variables:

County



- Linking with other datasets
 - Unemployment
 - Urban population density
 - Rural population density
- Running aggregates

Voter



- Committee IDs and keys associated with
 - County
 - State
 - Election Year

ETL



- FEC dataframe filtered using Colab to write and perform scripts
 - 10 columns lacking relevance removed

Donations



- Committee IDs mapped to Democratic or Republican in order to filter data
- Mapping from city to county level views
- Processed for each county in 6 states

Phase 1: Data Exploration

Before formulating our predictive framework, we needed to establish the following in order to glean any interesting patterns or trends, and if we could predict the results of an election:

**An event with
measurable output
(political election
outcomes in counties
of battleground
states)**

**An event having high
socio-economic value
with data available
for prediction
(donations data,
demographics,
unemployment
information)**



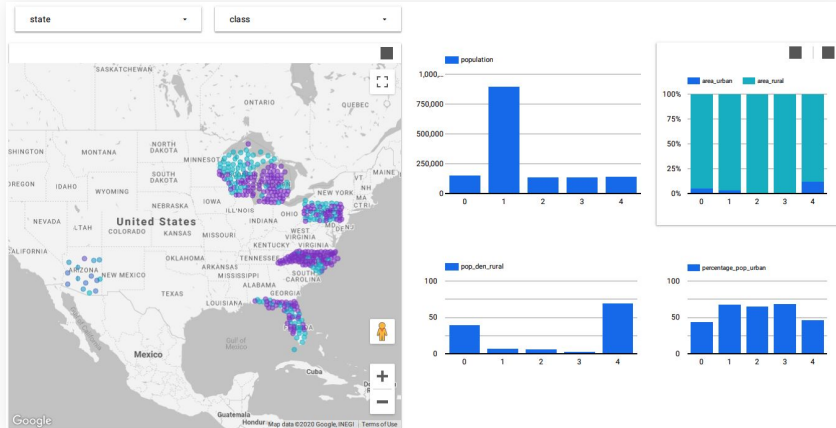
After establishing that we would use political donations data, we focused on building our model to answer the following questions:

- ① Can we predict who the president will be based on donation data?
- ① Are donations affected by unemployment?
- ① Do battleground states have a significant impact on elections?
- ① How do major socio-economic crises impact donations, and therefore votes?
- ① Do demographics, such as age or education, have an effect on whether voters choose to make political donations?

Phase 2: Data Analysis

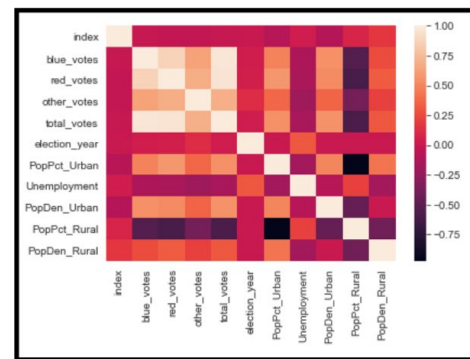
By performing preliminary data analysis, we made sense of the raw data, and determined the best data storage options and methods to pull the data for our machine learning model and predictions. Using several resources to guide our exploratory data analysis, we created the following visualizations and tables:

Donor Analysis



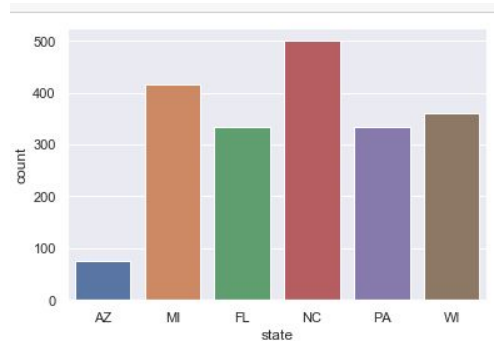
Sources for Analysis: + Medium Blogs; + [15 Data Exploration Techniques](#); + [Data Preparation and Exploration](#); + [Sisense](#)

Correlation Heatmaps



Candidates by State - 141,375

Unique Value Count



Examples of votes by county and candidates by state; aggregated county votes by state - 2020

Phase 2: Data Analysis

Our team used a variety of tools and tables to analyze the data:



Spreadsheets:

Everyone on the team participated in data gathering and located information to assist in our project. This information was located online, using data from the FEC, Kaggle and the Census Bureau. These data files were analyzed and saved as CSV files, which were uploaded to GitHub, PostgreSQL PGAdmin and Google Cloud for easy access and data sharing.



Google Cloud:

We placed large datasets and bucketed information onto Google Cloud. This allowed us to perform the ETL and join tables.



Jupyter Notebook:

This has been our main form of data analysis. We read in the tables from our SQL and Cloud connections to write simple to complex coding, focused on understanding and displaying the necessary data. To measure presidential data, we've pulled in donation and voter information, along with unemployment, education and demographics. We have discovered that these pieces of information have been useful in helping us transform and mold the data into information that is useful for looking at presidential elections.



PG Admin:

We use PGAdmin to perform SQL queries and analyze datasets, as well as review table details.

Phase 3: Machine Model

Our predictive model performs a linear and logistic regression analysis, using both supervised and unsupervised machine learning. The choice to use such models was informed by research and advice from our advisors.



Regression Analysis

Logistic: 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.



Regression Analysis

Linear: As this is a predictive method for continuous variables, we performed a linear regression in 2 parts:

- On aggregate number of votes based on donations
- Donations based on other non-binary features such as city, state, zip code and employer



Why not a neural network model?

- 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 models such as RandomForest and regressions.



How is our training and testing data split?

- Based on what is considered 'industry standard' in data visualization, our training and testing data is split in 75% training, and 25% testing.

Phase 3: Machine Model



Regression Analysis Considerations

Logistic: Advantages

- Trained weights give inference about the importance of each feature. Logistic regression can be used to find the relationship between the features.
- Updates to the model can be done easily unlike SVM or decision trees.
- Outputs well calibrated to probabilities. It is less prone to overfitting in a low dimensional dataset.

Logistic: Disadvantages

- On high dimensional datasets, may lead to over fitting since it attempts to predict precise probabilistic outcomes based on independent features.
- Non linear problems can't be solved since it has a linear decision surface.
- Difficult to capture complex relationships
- In Linear Regression independent and dependent variables should be related linearly. But Logistic Regression requires that independent variables are linearly related to the log odds ($\log(p/(1-p))$).
- Sensitive to outliers



Regression Analysis Considerations

Linear: Advantages

- Easy to implement and interpret the coefficients, and thereby understanding the relationship between the independent and dependent variables.

Linear: Disadvantages

- Susceptible to overfitting but can be reduced by reduction techniques, regularization (L1, L2).
- Outliers can have large impacts on the regression and boundaries. Linear Regression assumes that there is a linear relationship between the two variables, it assumes independence between attributes. Linear regression looks at the relationship between the mean of the dependent variable and the independent variable, thus just as mean is not a complete description of the entire variable, nor is linear regression.



RandomForest Considerations

Advantages

- Decorrelates trees, gives each tree a subset of features.
- Reduced error, as it is an ensemble of decision trees to predict the outcome of a particular event. Good performance on imbalanced datasets.
- Can work with big data
- Can handle missing data
- Little impact from outliers
- No problem of overfitting

Disadvantages

- Appears as a black box
- Features need to have some predictive power or they won't work
- Predictions of the trees need to be uncorrelated



K-NN (Nearest Neighbors)

Unsupervised ML

Advantages

- Can detect what is non-detectable to the human
- Can discover potentially powerful hidden patterns with data
- Exploratory data analysis can be run to understand process

Disadvantages

- More costly, more difficult than supervised machine learning
- Difficult to interpret the results for value as the answers all lack labels
- Does not work well on datasets with large dimensionality
- Scaling is a must
- Sensitive to outliers and can't handle missing values



<text>

<text>



<text>

<text>



<text>

<text>

Results Of Modeling

<text>

Dashboard Storyboard

Model Summary

← → 127.0.0.1:5000 ☆ Incognito

Biden || Trump?

Which Presidential Candidate will win the Swing States in 2020?

Can we predict the number of votes obtained based on the amount of donations raised in that county?

Team5k Home Page 2 Team

Linear Regression

Unsupervised

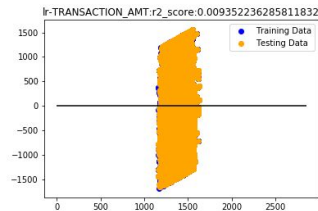
Stats: Donations

Stats: Votes

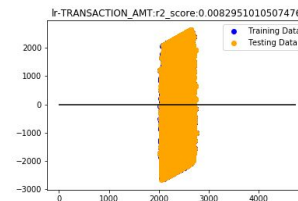
Submit

Select model type: Logistic or Linear

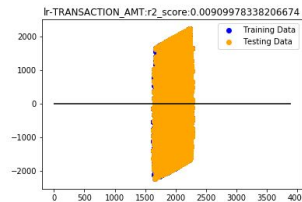
Arizona



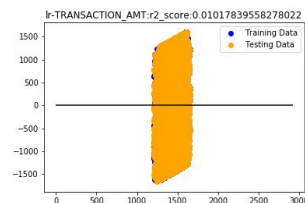
Florida



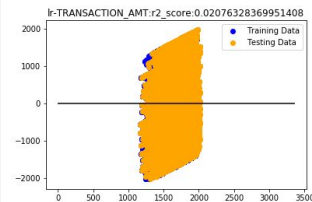
Pennsylvania



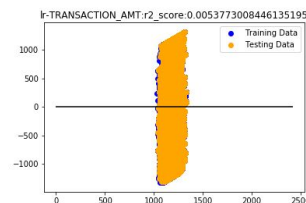
North Carolina



Michigan



Wisconsin



Dashboard Storyboard

Supervised Model Summary

127.0.0.1:5000 Incognito

Biden || Trump?

Which Presidential Candidate will win the Swing States in 2020?

Can we predict the number of votes obtained based on the amount of donations raised in that county?

Team5K Home Page 2 Team

Unorganized
Stats: Donations
Stats: Votes

Submit

Select State

AZ
FL
NC
PA
MI
WI

Select Model

Logistic
Linear

Model Run

X
Y
Z

Confusion Matrix and
Accuracy Scores

Model Plot

Agg Metrics

Dashboard Storyboard

Unsupervised Model Summary

← → 127.0.0.1:5000 ☆ Incognito

Biden || Trump?

Which Presidential Candidate will win the Swing States in 2020?

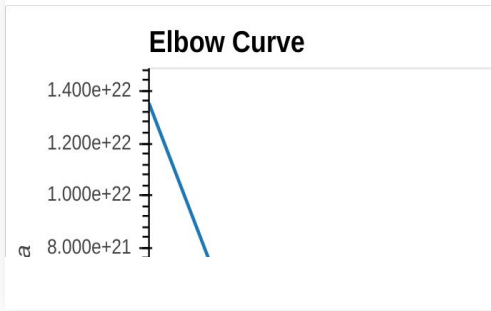
Can we predict the number of votes obtained based on the amount of donations raised in that county?

Team5K Home Page 2 Team

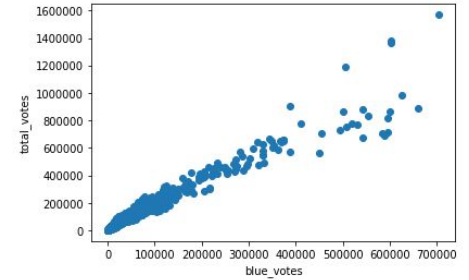
Unsupervised
Stats: Donations
Stats: Votes

Submit

Elbow Graph



Scatter Plot



Class A Summary



Class B Summary



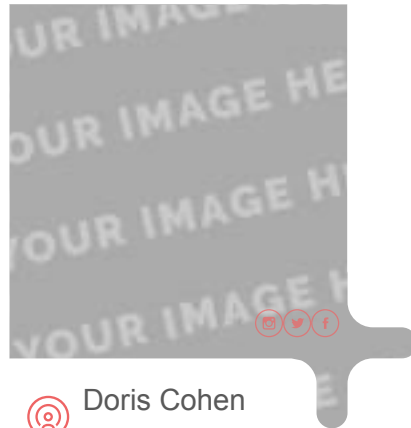
Class C Summary



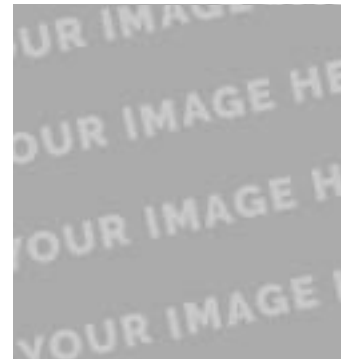
meet the team




 name here
job positions



 Doris Cohen
job positions



 name here
job positions



 name here
job positions

