# <Title>

## Introduction

This analysis sheds light on the predictive power of political donations on presidential election outcomes in key battleground states in the United States. In order to do so, we built predictive models using machine learning to understand how donations to political campaigns inform voting in counties and states, based on datasets collected from publicly available resources. Considering that we look at data between 2000 and 2020, we also assess 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.

We used both a logistic regression model paired with an unsupervised clustering model in order to build our predictive modeling. <Insert short summary of why each model was used>

Our analysis indicates that…<Insert summary of results>

This analysis is split into a number of sections. First, we provide an overview of the questions we asked in order to perform our analysis. In the second section, we spend time describing the analytical processes involved, 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.

## Overview of Problem Statements

*Purpose of this section: Describe the problem. What substantive question are we trying to address?*

Our model captures the predictive power of political donations on presidential election outcomes. We also assess how crises between 2000 and 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?

## Data Analysis

*Purpose of this section: What data did you use to address the question, and how did you do it? When describing your approach, be specific. For example: Don’t say, “I ran a regression” when you instead can say, “I fit a linear regression model to predict price that included a house’s size and neighborhood as predictors.” Justify important features of your modeling approach. For example: “Neighborhood was included as a categorical predictor in the model because Figure 2 indicated clear differences in price across the neighborhoods.”*

In order to prepare to build our model, we first began by performing a standard data analysis. This included steps to clarify data requirement specification, data collection, data processing, data cleaning, and finally, our data analysis.

***Data Requirements Specification***

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.

|  |  |  |
| --- | --- | --- |
|  | Features | Sources |
| Voting Data | Election outcomes during various crisis years (Dotcom bubble burst, financial crisis, COVID19) | [Electoral College Votes](https://data.world/government/us-election-results), [Votes by State](https://worldpopulationreview.com/state-rankings/electoral-votes-by-state) |
| Demographics | Urban/Suburban split, ethnicity, county information, zipcodes | [Health Metrics](http://ghdx.healthdata.org/gbd-results-tool), [Census Bureau Data](https://www.census.gov/) |
| Unemployment Data | Year, month, season | [Census Bureau Data](https://www.census.gov/programs-surveys/popest/guidance.html) |
| Donations Data | Split for 6 states, county-specific donations, bipartisan splits | [FEC Campaign Finance Data](https://www.fec.gov/campaign-finance-data/party-code-descriptions/), [FEC General Data](https://www.fec.gov/data/browse-data/?tab=bulk-data) |

<Insert discussion of split between numerical and categorical data>

***Data Collection***

*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 dataframe 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.

***Data Cleaning***

*Logistic Regression Model:* Key features in this model come from the Donor file for the six swing states. We selected the Committee City, State, Zip, the transaction amount, and donor's employer, and donor's occupation. We performed the following joins:

* County
  + Linking with other datasets
    - Unemployment
    - Urban population density
    - Rural population density
  + Running aggregate
* Donations
  + Pull 4-year segment
  + Processed
  + Summation
  + Went from city level to county view
  + Processed for every county, within 6 states
  + Committee ID identifies with party the committee donates to. Mapping the committee IDs to Republican or Democratic. This is what we’re using to filter all the donations into either red or blue.
* Voter
  + Keying donations from County, State and Election Year
* ETL
  + FEC dataframe was filtered using Colab to write and perform the scripts
    - Removed approximately 10 columns that were not relevant
    - Kept items that were needed for the analysis

*Unsupervised Clustering:* All metrics were placed into a PCA analysis. The PCA analysis showed that only 3 features would really drive segmentation. From there, an analysis of model inertia shows that we could select between 3 and 5 clusters. We took 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 what we needed to remove. Then, we sorted thru the data and optimized it for our project. We believe, from the data, that looking at previous voter information, coupled with donation info and demographics might provide insights on our focus which is donations affecting presidential elections in swing states. We’re testing to see if there is a correlation between the money people donate and who is elected. We performed the following joins:

* County
  + Linking with other datasets
    - Unemployment
    - Urban population density
    - Rural population density
  + Running aggregate
* Donations
  + Pull 4-year segment
  + Processed
  + Summation
  + Went from city level to county view
  + Processed for every county, within 6 states
  + Committee ID identifies with party the committee donates to. Mapping the committee IDs to Republican or Democratic. This is what we’re using to filter all the donations into either red or blue.
* Voter
  + Keying donations from County, State and Election Year
* ETL
  + FEC dataframe was filtered using Colab to write and perform the scripts
    - Removed approximately 10 columns that were not relevant
    - Kept items that were needed for the analysis

## Model Summary

**Logistic Regression**

The goal of the logistic regression is to determine based upon key attributes of donors to determine if the party is Republican or Democrat

**Training / Testing Split**

The typical split from training and testing was used. Given that there were similar amounts of Republican and Democrat donors, there was no need to over or under sample.

**Model Choice**

The logistic regression model was used because a donor is either Democratic or Republican, so the binary choice was appropriate. We also compared the Logistic Regression with a Random Forest.

**Model Accuracy**

The logistic regression model was fairly accurate with a score of 91.8%. The Random forest had a predictive accuracy rate of 99.9%, which may indicate overfitting

**Model Success**

This model will help candidates target donors by accurately predicting which party that they will donate.

**Unsupervised Clustering**

The goal of the clustering model is to segment individual counties based upon a range of unemployment and health metrics. A separate model was created for each metric

**Training / Testing Split**

Not applicable for unsupervised learning

**Model Choice**

The point of this model is to help campaigns create locally targeted messaging to speak to donors based upon population, unemployment, and health metrics

**Model Success**

Placing this model in the dashboard provided an easy view to help campaigns identify trends and clusters by county. <Can we include a discussion here about some metrics we are using to measure success for the model?>

## Results and Practical Applications

<Insert insights from results, paired with dashboard PNGs and PNGs collected from model analysis>

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.