Final Project Presentation

Thomas Lamont & Daniel Rivera

Dec 2024

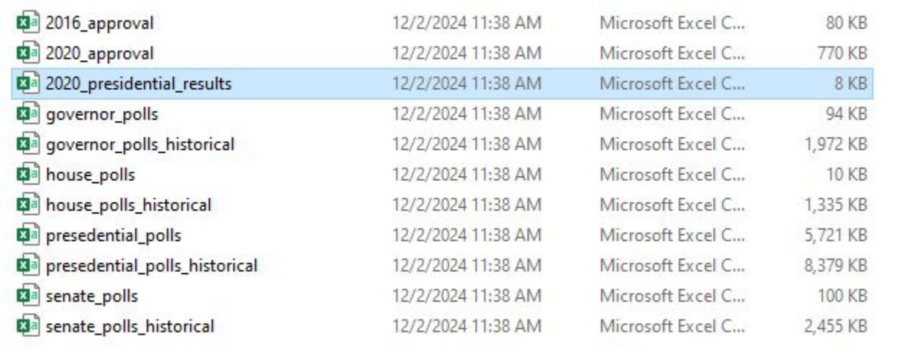
Overview

- Goal: To accurately predict the 2024 U.S. presidential election results with use of machine learning models.
- Data Source: Polls, Election results, approval ratings

Machine Learning Life Cycle

1. Data Collection
   - We used 2 sources a huge overall dataset of polls and approval rates, among other things. Another for specifically 2020 election results by state.
2. Data Preparation
   - Most of our time was spent on this part. Removing a lot of noise and addressing missing data. As well as, combining specific files for easier use.
3. Model Selection and Training
   - Applying Logistic Regression, Random Forest, Gradient Boosting to our training data.
4. Exploratory Data Analysis and Parameter Tuning
   - Approval Ratings over time
   - Pre Election Consensus
5. Model Evaluation and Testing
   - State Prediction Visualization
   - Our model prediction Correctness
     i. Correctly predicting the swing states
6. Prediction and Interpretation
   - Using our trained models to predict the 2024 election outcome

Chosen Data



2016_approval	12/2/2024 11:38 AM	Microsoft Excel C...	80 KB
2020_approval	12/2/2024 11:38 AM	Microsoft Excel C...	770 KB
2020_presidential_results	12/2/2024 11:38 AM	Microsoft Excel C...	8 KB
governor_polls	12/2/2024 11:38 AM	Microsoft Excel C...	94 KB
governor_polls_historical	12/2/2024 11:38 AM	Microsoft Excel C...	1,972 KB
house_polls	12/2/2024 11:38 AM	Microsoft Excel C...	10 KB
house_polls_historical	12/2/2024 11:38 AM	Microsoft Excel C...	1,335 KB
presedential_polls	12/2/2024 11:38 AM	Microsoft Excel C...	5,721 KB
presedential_polls_historical	12/2/2024 11:38 AM	Microsoft Excel C...	8,379 KB
senate_polls	12/2/2024 11:38 AM	Microsoft Excel C...	100 KB
senate_polls_historical	12/2/2024 11:38 AM	Microsoft Excel C...	2,455 KB

- Combines polling data and 2020 election results
- Assigns the outcome labels
- Feature Engineering
  - Merged polls with historical data (swing state indicators/pollster grades)
- Deeper understanding of polling, electoral systems, and historical data

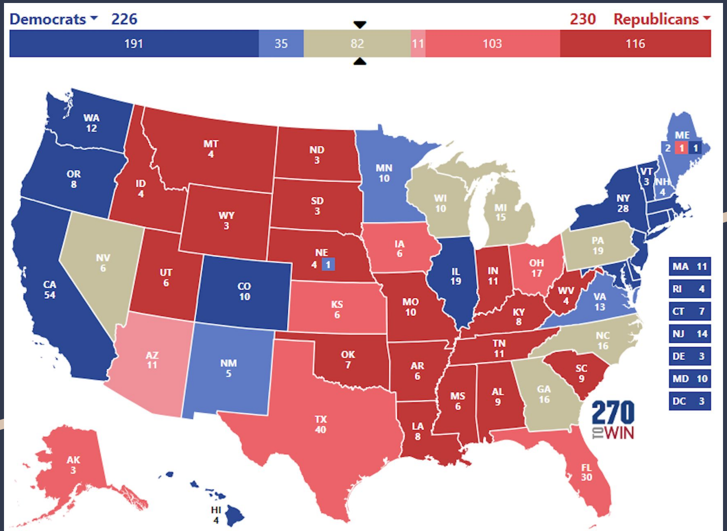
Cleaned and condensed data

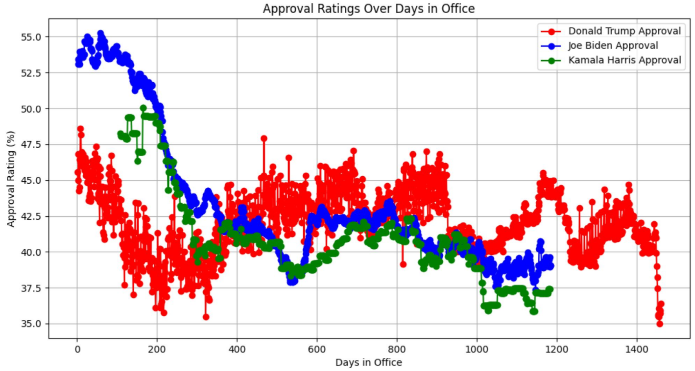
- Started with 59113, 52 from polling data
  - dropped unneeded/noisy columns
  - Added poll type for house/senate etc
  - Rows for each candidate condensed
  - Classified 2020 results
    - Safe/Likely/Leans
  - Pollster grading inconsistent over files

- final shape 19984, 38 in a single file

Exploratory Data Analysis

- Critical Predictors
  - Approval Ratings
  - Swing States
- Historically these 2 factors were (and still are) extremely big factors in determining elections





Model Selection

- Random Forest
- Gradient Boosting
- Logistic Regression

- Our projects dataset involves extremely complex sources where 1 file can contain up to 50 columns of numerical and categorical data. RF can model high-dimensional data which helps the variety of features used.
- Elections have critical “hard-to-predict” states/scenarios (swing states), and GB can be extremely effective in these nuanced cases.
- Logistic Regression is a familiar baseline model in comparison to the earlier mentioned models. We used this to get started since something like high approval ratings correlated with wins can be much more simple to predict.

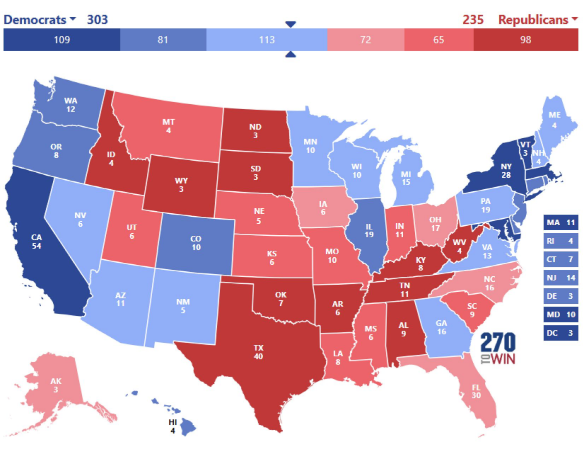
Results...........

(drum roll please)

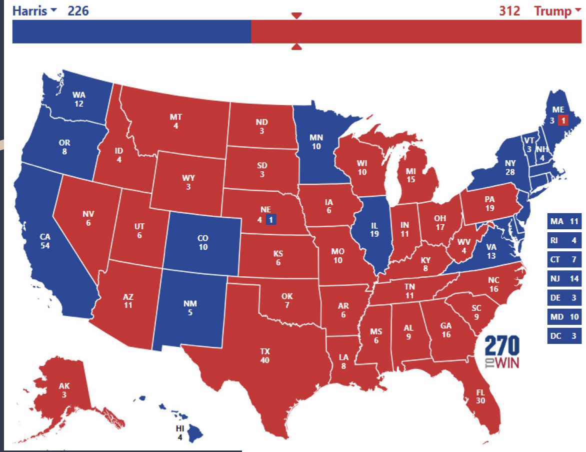
Kamala!

Our model indicates that Kamala and Waltz will win the 2024 election. Swing States played a huge role in this prediction which our model unfortunately failed at predicting.

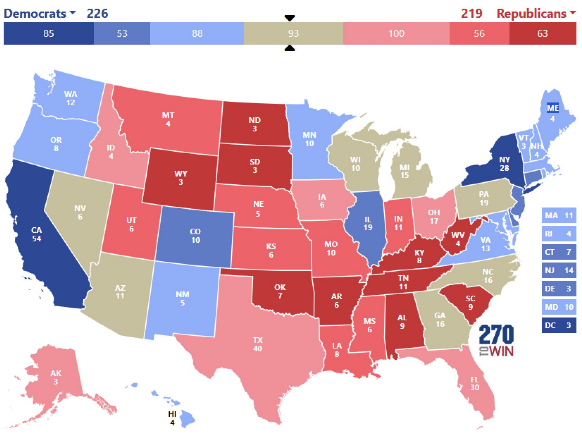
Prediction



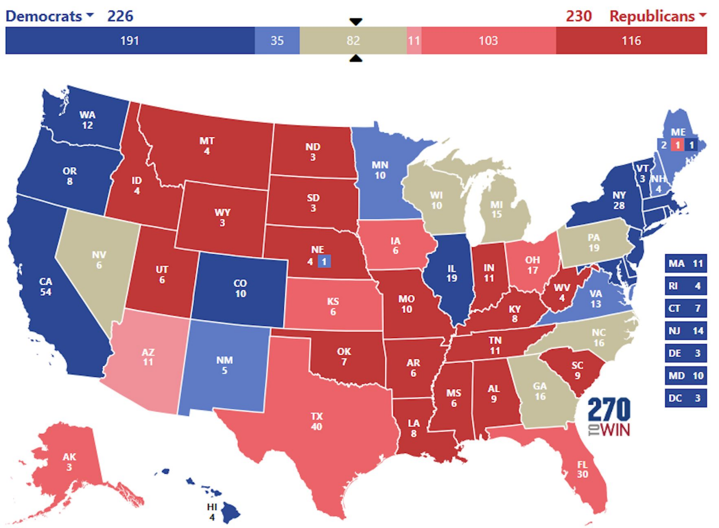
Actual



No Swing States



Challenges



- Some states 1 party consistently wins over the other, which can can underperformance in swing states
  - Consensus map is much more sure than our predictions
- Unpredictable events
  - Elons full support of Trump
  - The Assination attempt
  - Kamala’s list of Celebrities
- Overfitting due to irrelevant historical events
  - 2008 Great Recession
  - COVID-19 in the 2020 election

Findings:

- Machine learning models can predict election outcomes with reasonable accuracy.
- State-level data significantly improves results.

Future Work:

- Incorporating more granular polling data.
- Real-time predictions during election cycles.

Datasets used

https://www.kaggle.com/datasets/jpmiller/elections/data

The same website we showed previously in our demo

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/42MVDX

2020 Election by state Data which was used for the classification

Questions?