Classification and Clustering Methods to Predict Pollster Accuracy in US Elections

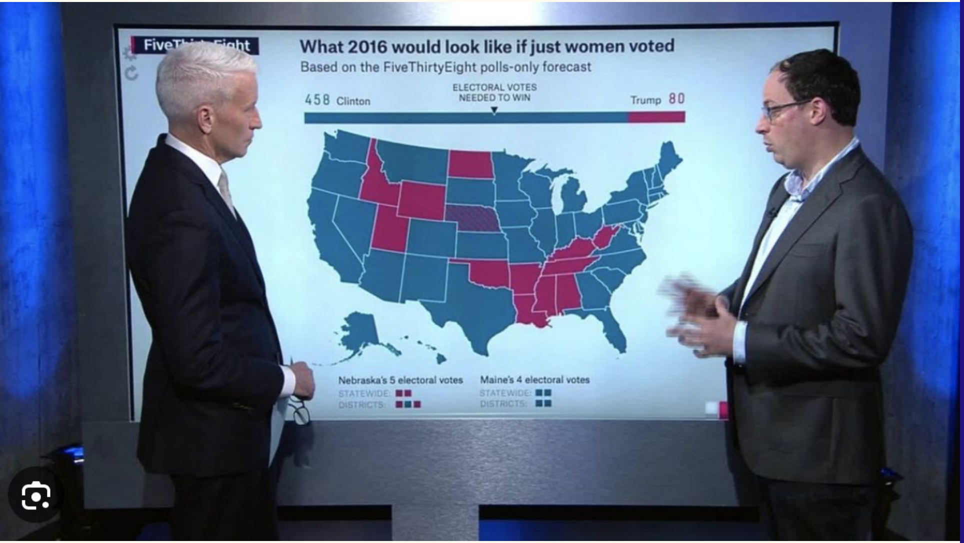
Vijay Ravuri
Nikki Hassell

Problem Statement

We will investigate polling accuracy across different methodologies, partisan status, and type of race. Our goal is to predict if a poll will make the right call based on polling features or discover similar trends in accurate or inaccurate polls based on clustering methods.

The Data

FiveThirtyEight



Picture Source: CNN
Picture Source: Nieman Lab

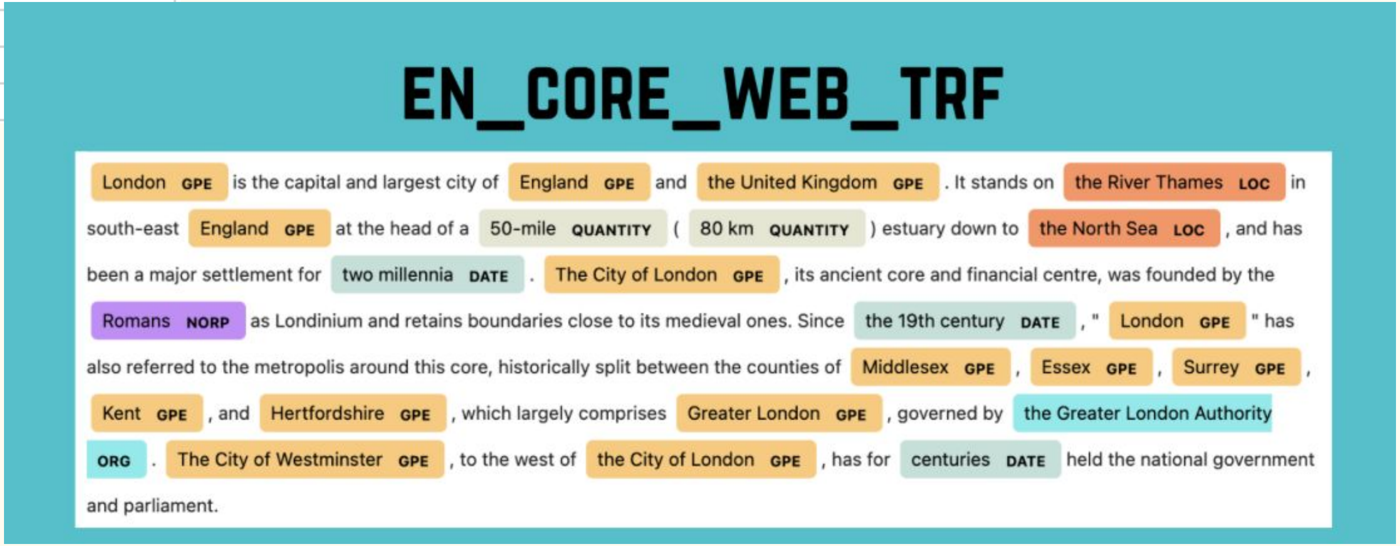
- Started in 2008 as a blog
- Namesake - 538 electors in US electoral college
- 2013 - acquired by ESPN
- 2018 - transferred to ABC
- Broad spectrum of subjects now focusing on elections, politics, and American society

Using SpaCy Models for Comment Detail

comment
for New York Daily News | WABC-TV (New York)
for unspecified Democratic sponsor
for New York Daily News | WABC-TV (New York)
for Charles E. Schumer
for Richard Stallings
for Richard Stallings
sample size unavailable; estimated at 600 as a default
for Tom Vilsack
for Richard A. Hill
for Robert C. Hayes
for CNN | Time

Flags for **organization**, **person**, poll for a **unspecified D/R sponsor**

Regex to flag: among registered voters, average of multiple versions/turnout models listed, and imputed sample size



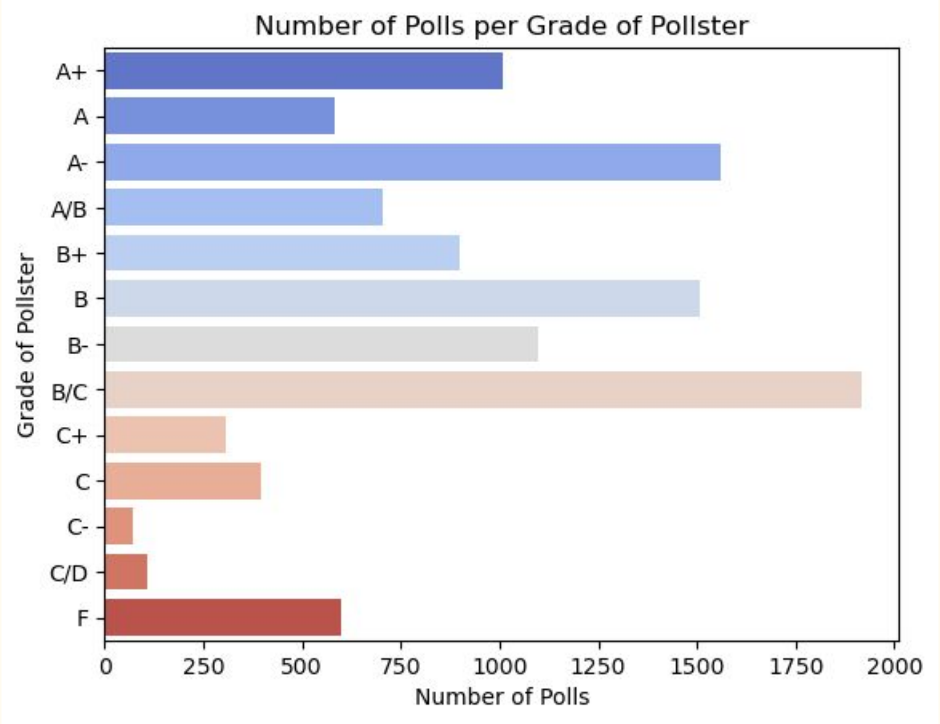
Sources: [TRF image]
[SpaCy]

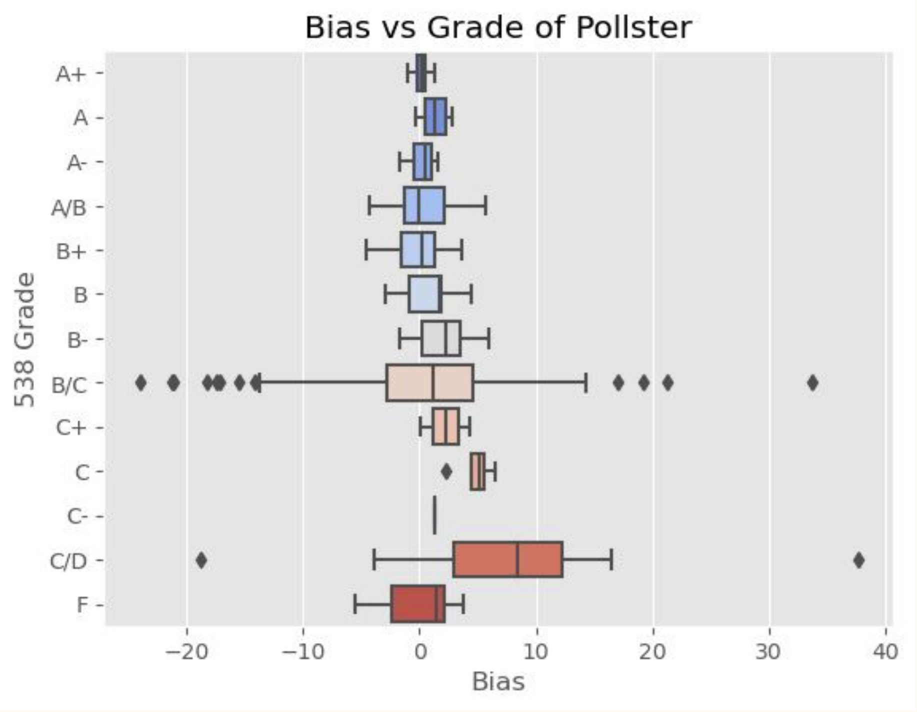
Exploratory Data Analysis

EDA on Pollster Grades

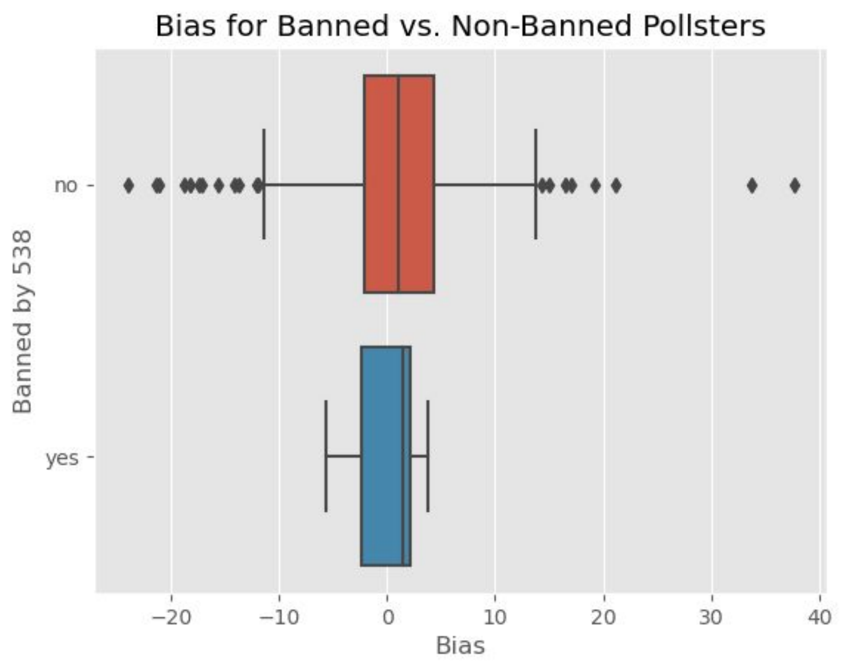
Grade	# of Pollsters
B- or Higher	98
B/C	382
C+ or Lower	37

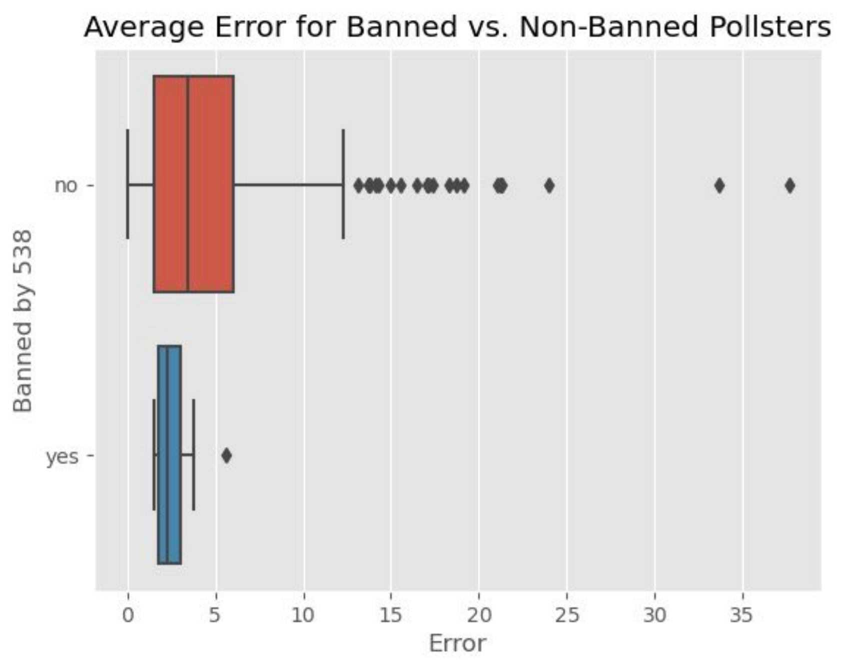
Context on 538 Grading: 538



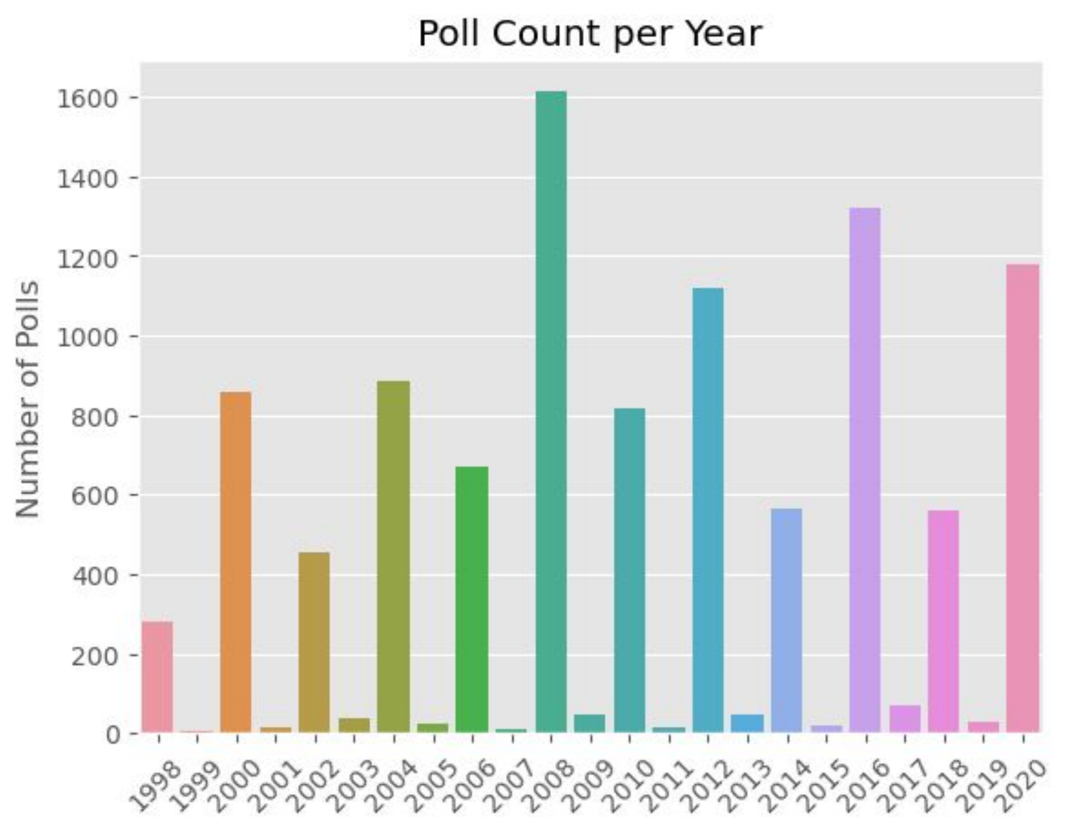


EDA on Pollster Grades





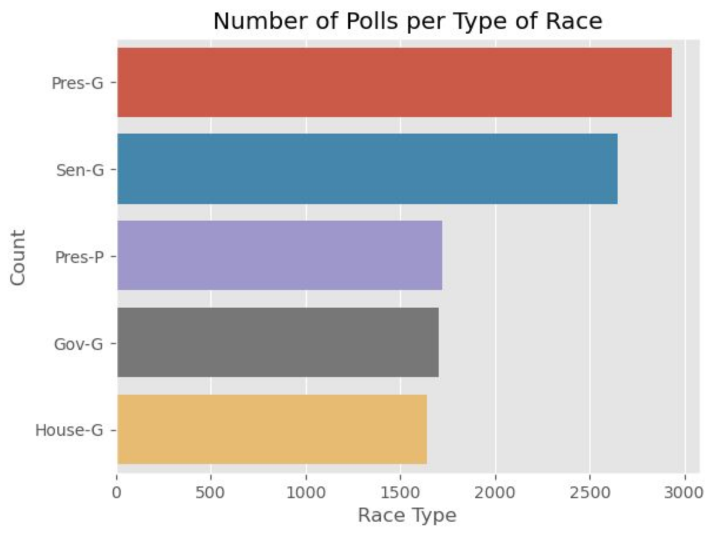
Number of Polls - Spike in Polls in 2008



Increase in polls in **1998-2008**

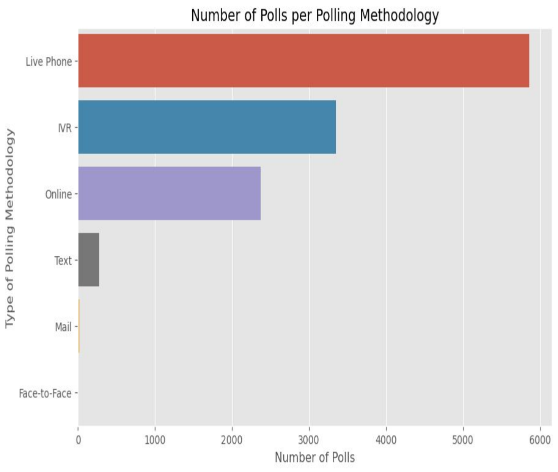
Access to **online** and **text** capabilities

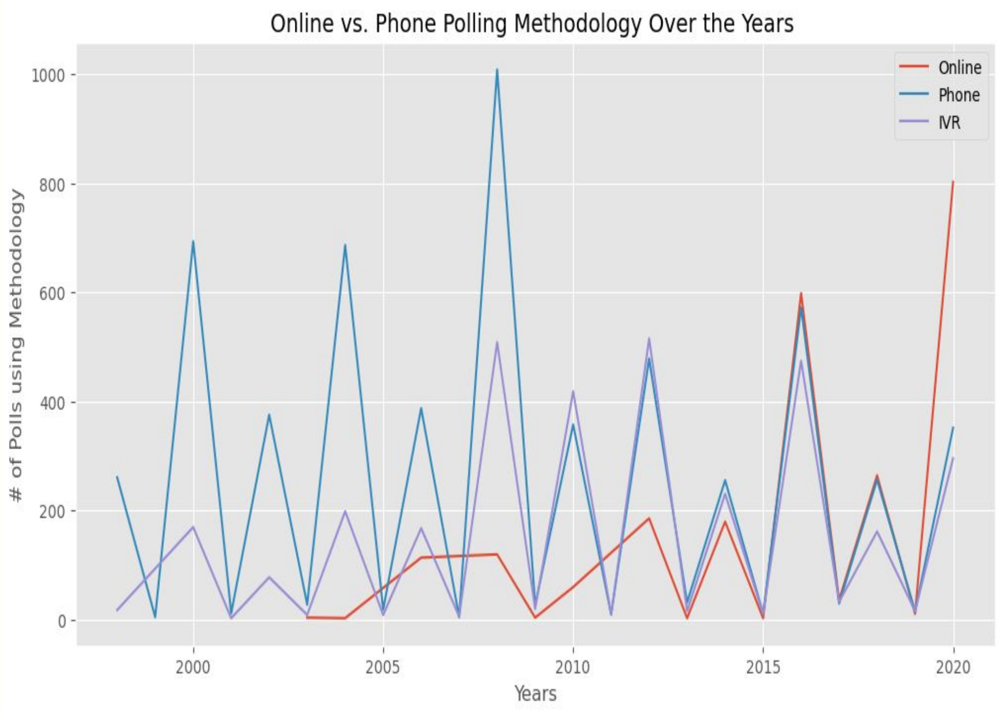
Most polling done on **Presidential Election**



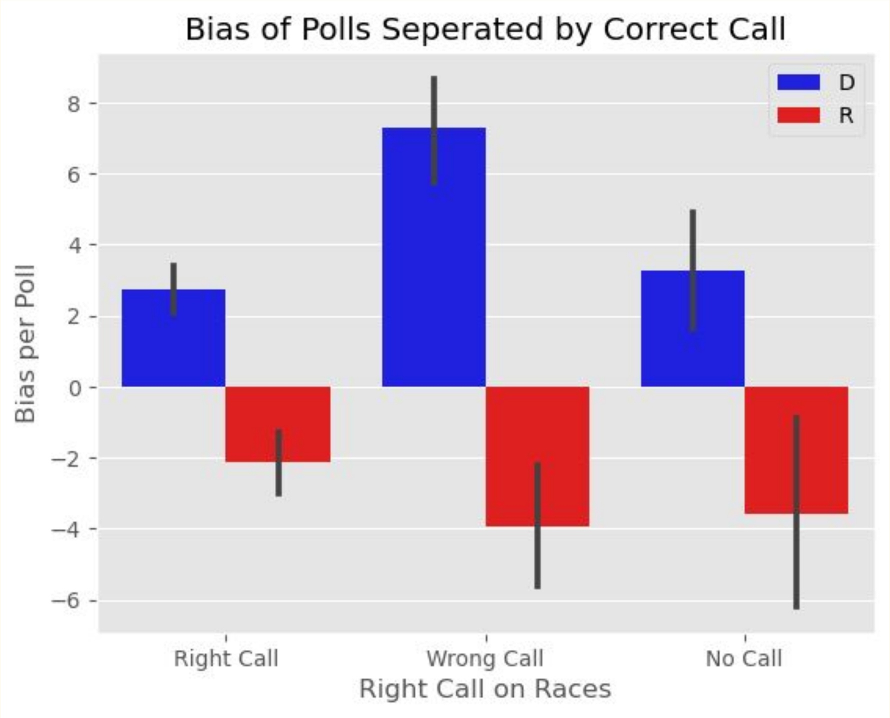
Types of Methodologies

Live Phone, IVR, and Online polls were by far the most common methodologies.





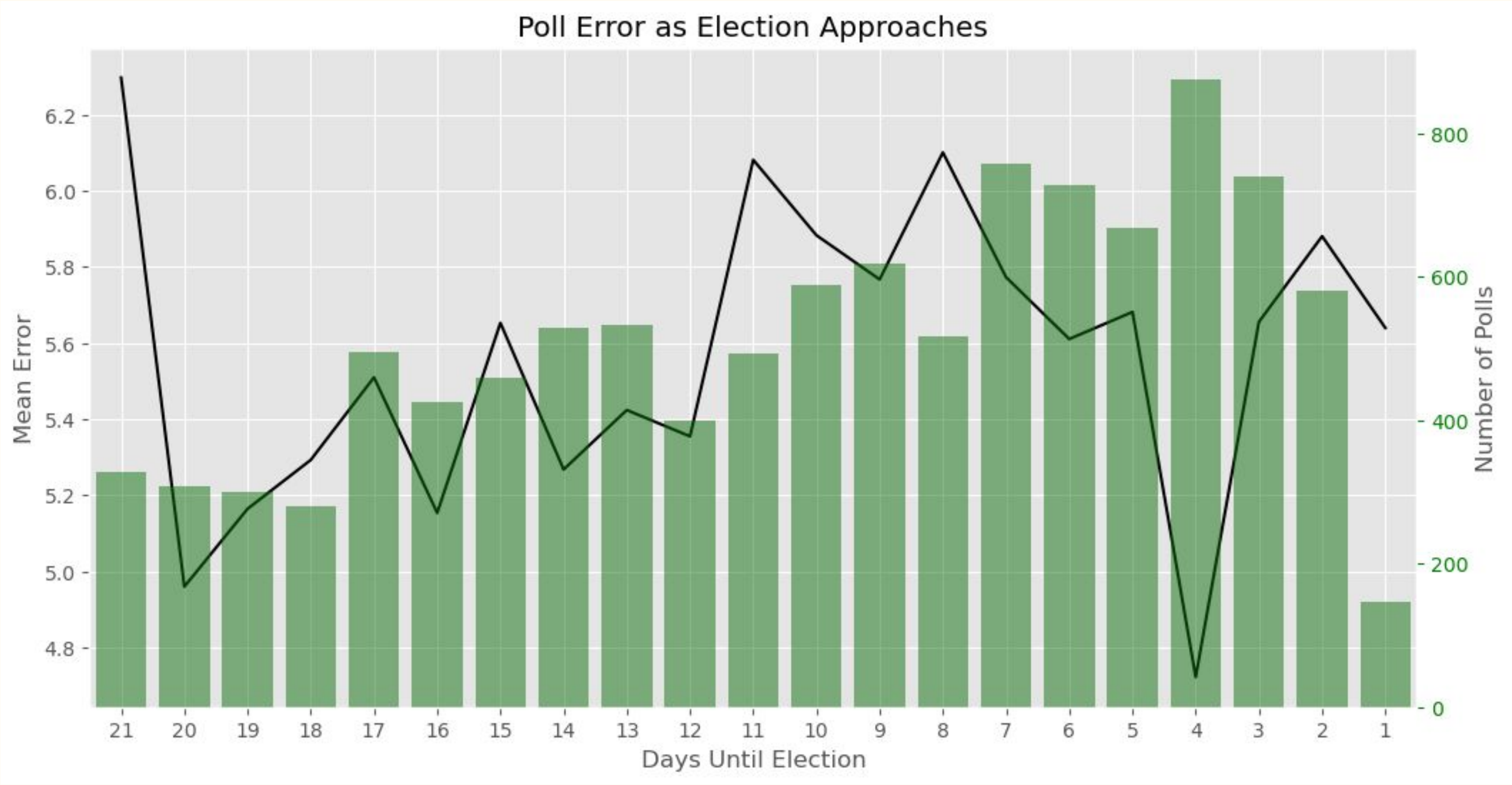
Understanding Bias vs. Error



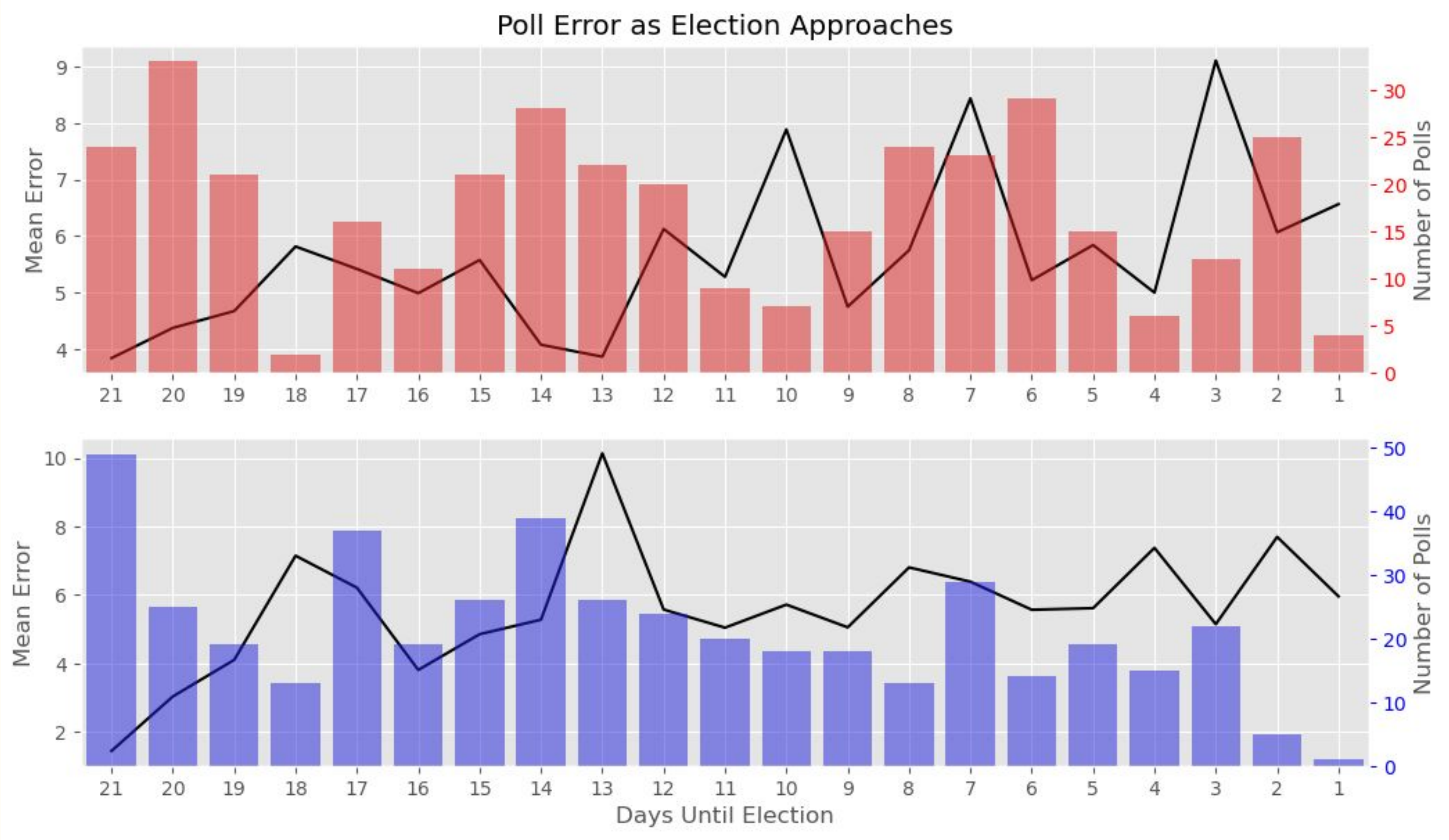
- **Positive Bias Score** - favorable to a Democratic candidate compared to actual score
- **Negative Bias Score** - favorable to Republican candidate compared to actual score
- **Error** - Absolute Value of Bias

	Wrong Call	No Call	Right Call
Democratic (D)	121	36	294
Republican (R)	98	19	244

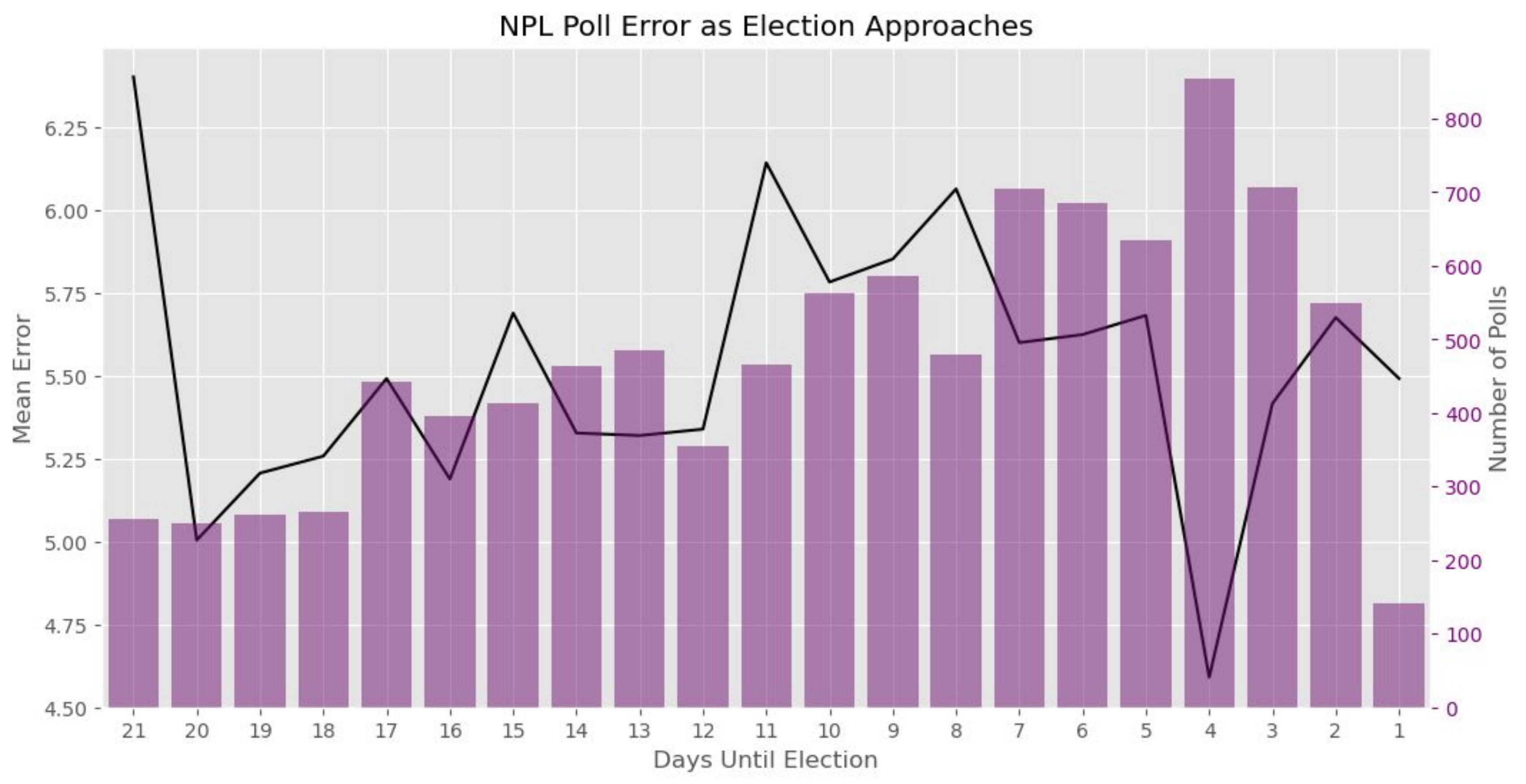
Poll Error as Election Approaches



Poll Error as Election Approaches by Party



Poll Error as Election Approaches - No Party Label



Prediction Methods

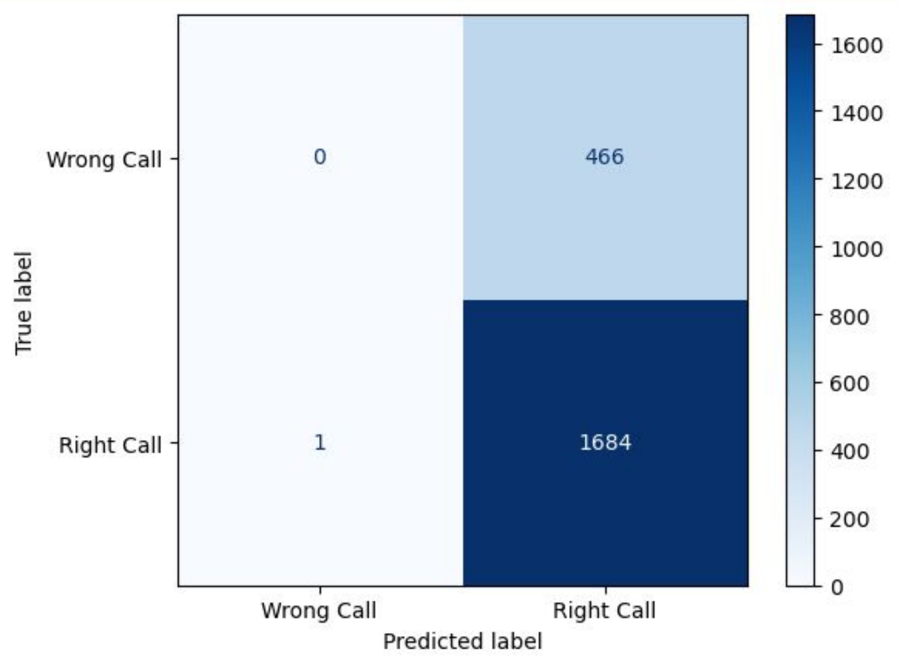
Modeling and Analysis:

Target - if a Poll made the Right call

1. Logistic Regression with Numeric Features (using StandardScaler and MinMaxScaler)
2. Logistic Regression + Categorical Features (using StandardScaler)
3. Simple Decision Tree (Cat + Num features)(using StandardScaler and MinMaxScaler) - Best Model
4. Grid Search Decision Tree - (Numeric and Categorical) (StandardScaler)
5. Random Forest Model - (Numeric and Categorical) (StandardScaler)
6. Random Forest with Interaction Features using PolynomialFeatures
7. Analyzing our best model (RF) Performance by Partisan group

Logistic Regression - Numeric Features

With StandardScaler

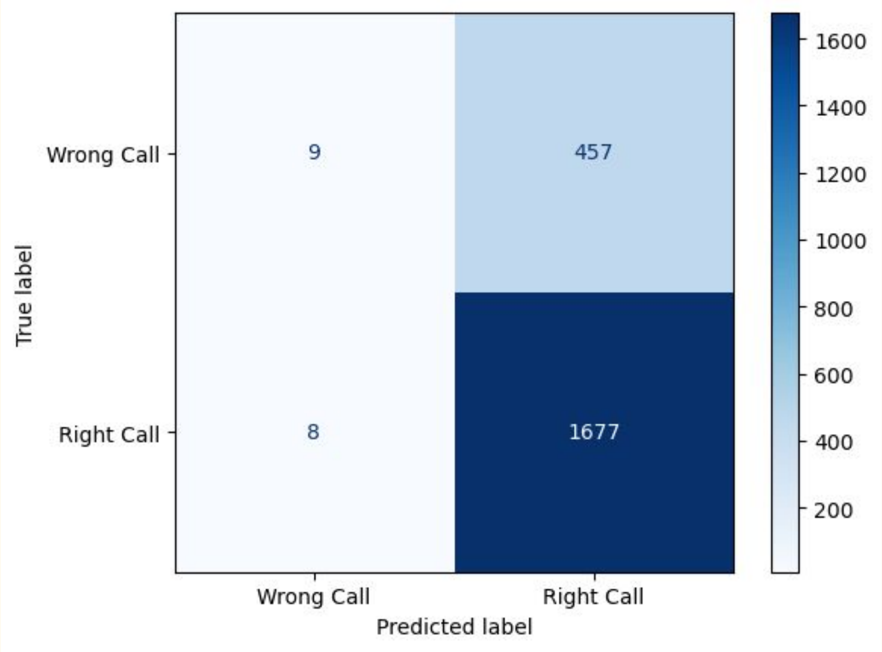


- **Numeric features** - 'year', 'samplesize', 'cand1_pct', 'cand2_pct', 'days_bt_polldate_election'
- **Target - Mapped Right Call** - 0 or 1 - 0.5 (dead heat imputed to 0)

Baseline	0.7835
Train Score	0.7827
Test Score	0.7829
Recall	0.9994
Precision	0.7833
f1 score	0.8782

Logistic Regression - Numeric and Categorical Features

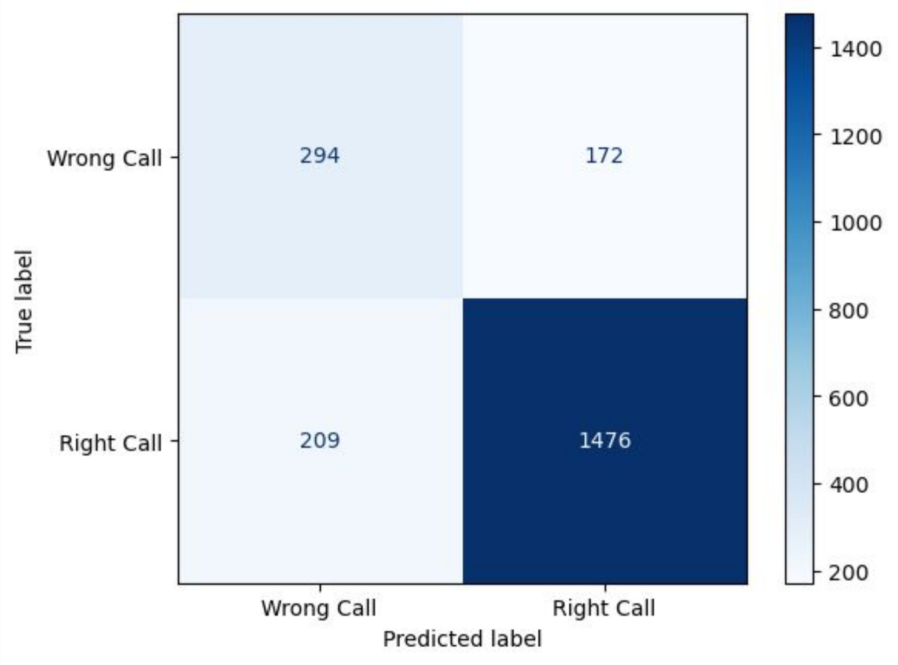
With StandardScaler



- **Numeric features**
- **Categorical** features - 'org', 'person', 'anon', ‘'registered_voters', 'averaged', 'imputed_600', 'Text', 'Live Phone', 'Mail',’ Face-to-Face', 'IVR', 'Online', '538 Grade'
- **Target - Mapped Right Call** - 0 or 1 - 0.5 (dead heat imputed to 0)

Baseline	0.7835
Train Score	0.7810
Test Score	0.7838
Recall	0.9953
Precision	0.7858
f1 score	0.8782

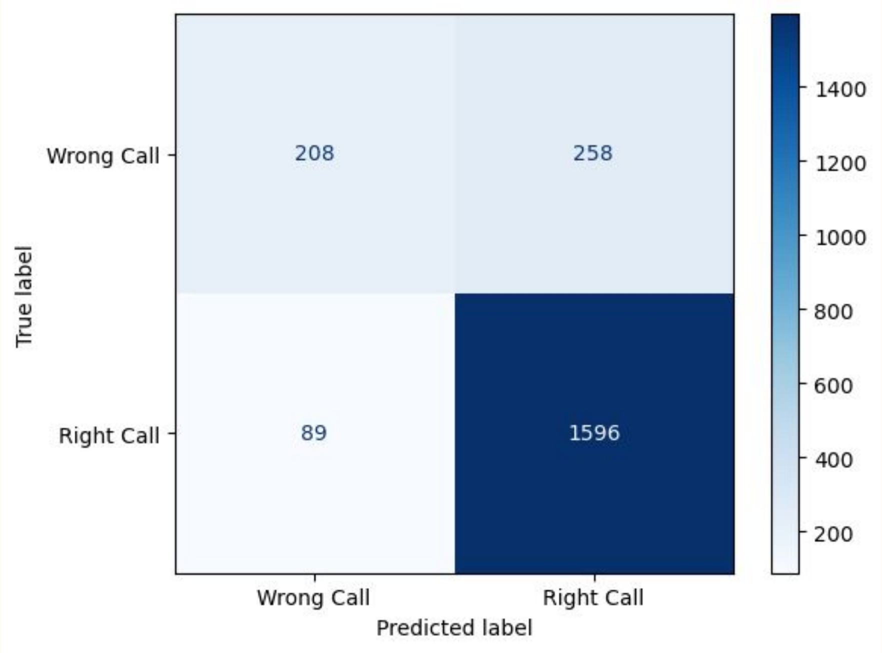
Decision Trees



- StandardScaler preformed pretty much the same MinMaxScaler
- Grid Search did not make much improvement compared to default

Baseline	0.7835
Train Score	0.8133
Test Score	0.8229
Recall	0.8760
Precision	0.8956
f1 score	0.8857

Random Forest Model

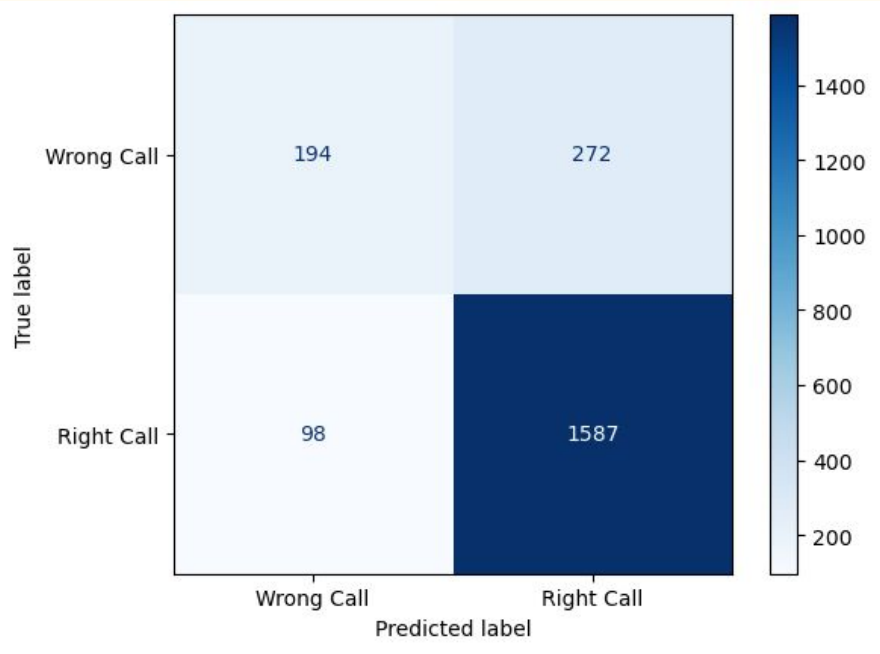


- **Numeric and Categorical features**
- **Target - Mapped Right Call** - 0 or 1 - 0.5 (dead heat imputed to 0)

Baseline	0.7835
Train Score	0.8332
Test Score	0.8387
Recall	0.9472
Precision	0.8608
f1 score	0.9019

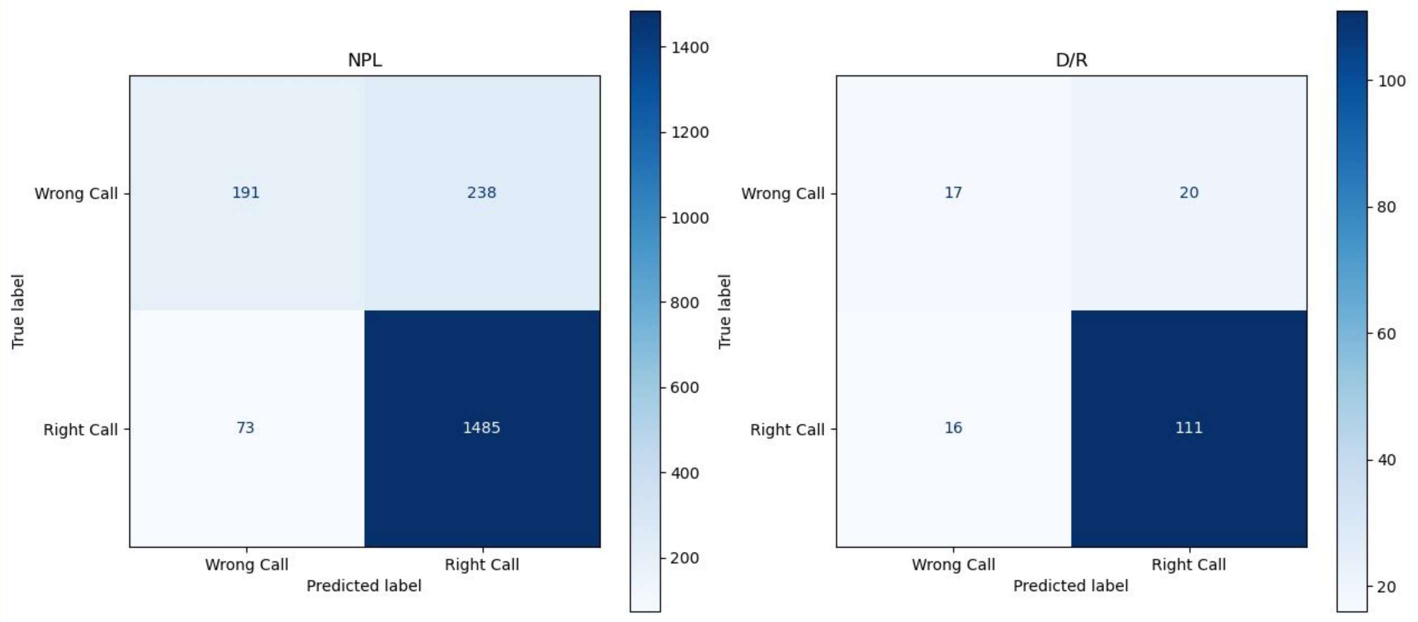
Random Forest with Interaction Features

- Used **PolynomialFeatures** on
  - Three most popular methodologies (Live Phone, IVR, Online)
  - Imputed 600 and sample size



Baseline	0.7835
Train Score	0.8269
Test Score	0.8280
Recall	0.9418
Precision	0.8537
f1 score	0.8956

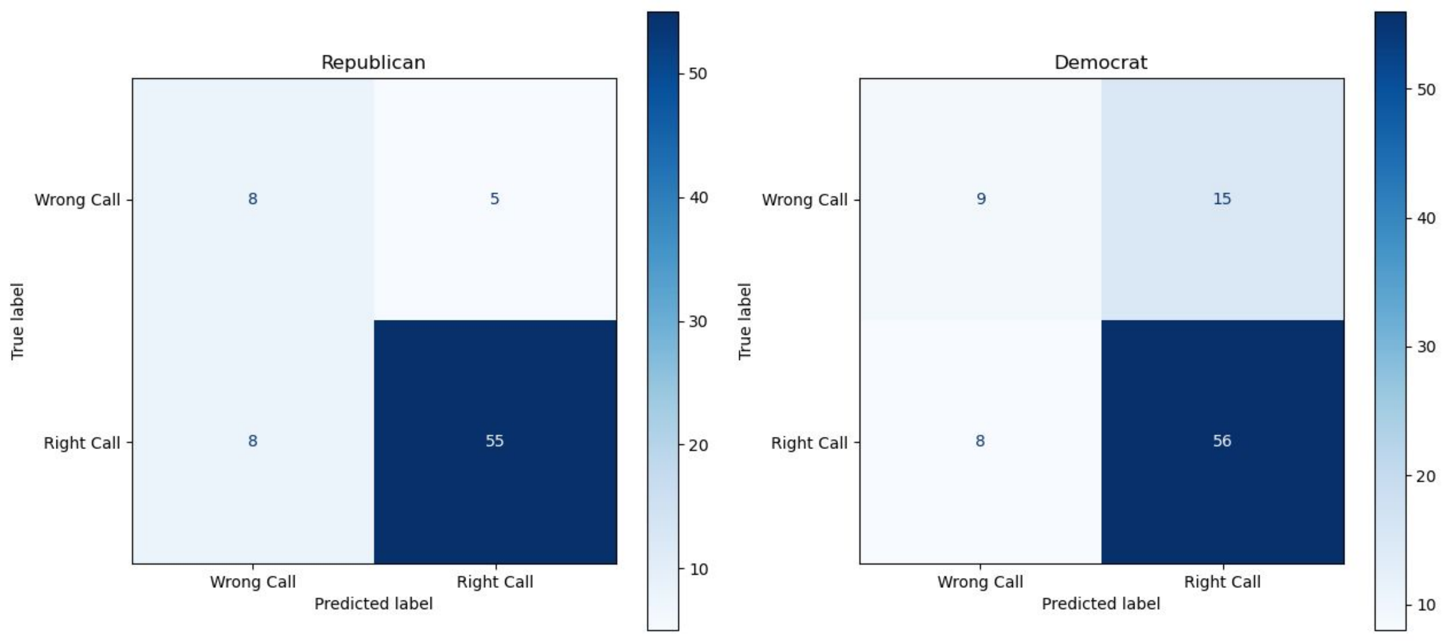
Analyzing our Best Model by Party - Random Forest



Partisan	# of Polls
No Party Label	9937
Democrat	451
Republican	367

	NPL	D/R
Accuracy	0.8435	0.7805
F1 Score	0.9052	0.8605

Analyzing our Best Model by Party - Random Forest



Partisan	# of Polls
No Party Label	9937
Democrat	451
Republican	367

	D	R
Accuracy	0.7386	0.8289
F1 Score	0.8296	0.8943

Clustering Methods

Process

- Three clustering algorithms attempted:
  - K-Means
  - DBSCAN
  - Hierarchical DBSCAN
- Compared results between Standard Scaler and Min-Max Scaler
  - Min-Max Scaler separates categorical features much more than Standard Scaler
  - Which scaling method was better depended on the clustering algorithm
- Applied Principle Component Analysis (PCA)
  - Used 85% of the variance as our cutoff
  - Did not change the results by much but did improve our clustering slightly
- Used Silhouette Score as our performance metric
  - Compares within cluster similarity with outside of cluster similarity(range: [-1,1], higher is better)
  - Not an appropriate measure for density-based clustering!

K-Means

- Min-Max Models:
  - Separated primarily on methodology
  - 2-means model split Live Phone polls into a cluster and the rest into the second
  - 3-mean model made separate clusters for Live Phone, Online, and IVR
- Standard Scaled Model:
  - Isolated some partisan polls that were a bit less accurate on average (~10% lower)
- Generally, K-Means did not show particularly interesting clusters

K	Scaler	Score
2	Min-Max	0.634
3	Min-Max	0.598
2	Standard	0.675

DBSCAN

- Standard Scaler gave very poor results
  - Standard Scaler best silhouette score: 0.095
    - Unable to find any meaning in the clusters
- Min-Max gave mixed but very interesting results
  - Best silhouette score: 0.423
  - Only created 1 cluster and identified 28 outliers
- The Outliers:
  - Very inaccurate polls (26.79% accuracy)
  - All partisan (15 Democratic, 13 Republican)
  - Ranged from 2012-2020
  - Almost all were done for unspecified donors

Hierarchical DBSCAN (HDBSCAN)

- Extension of DBSCAN that can detect clusters of varying densities
  - No more tuning epsilon!
- Identical best silhouette score for Min-Max and Standard scaler (0.383)
  - Completely different clusters however
  - Standard Scaler found clusters that did not seem meaningful
- Min-Max Scaler on the other hand:
  - Found 24 clusters + outliers
  - 8 of those clusters had 0 correct calls between them
    - 1070 incorrect calls (10% of our data)
    - 4 no-calls
  - All were nonpartisan polls
  - None were online polls
    - Likely coincidental

Conclusion

Conclusions

- Predicting poll accuracy through this approach is not easy
  - We were only able to improve our baseline by 5%
  - Polls are generally accurate and what causes them to be wrong is not easy to measure
    - Polls are trying to predict the future so there will always be some meaningful error present
- Clustering has some merit
  - Saw some interesting and meaningful clusters from all 3 approaches
  - Hard to tell whether clusters are valuable until we look at them very closely
- We did not find strong evidence of trends between partisan or methodological differences with error

Next Steps

Next Steps

- Using geopandas to create a heat map to visualize EDA by state
  - Methodology, Error, Bias, number of polls conducted
- Extend the clustering approach to find better separated clusters
  - Train a model to classify into those clusters to predict whether a poll is accurate by proxy
  - Implement a more effective measure of cluster quality
    - Density-Based Clustering Validation (DBCV)

Any Questions?