California Recall Prognostication for WAM! News

Wassie, Mekdes

Aziz, Zaid

Morgen, Nadia

WAM! News

Political

Problem Statement

Californians recently voted decisively not to recall their Governor. Yet many pollsters and pundits predicted a squeaker. An August SurveyUSA poll predicted that 51% of likely voters would vote to recall the Governor.

How could they be so wrong?

Problem Statement

- Why did the results differ so much from the prognostications?
- Is modern polling plagued by the lack of land lines that made random sampling easier?
- Did pollsters use the wrong metrics or miss some key metrics?

Goal

At WAM! News, we are trying to develop a more accurate model than our competitors.

Looking at 538's approach and Gustavo Caffaro's neural net approach as starting points

- Both predicted 2016 presidential better than most

Goal Metrics

Maximize:

- Accuracy
- Specificity
- Recall

Outperform our competitors… **WAM!**

Data

Cooperative Election Study (CES), fka Cooperative Congressional Election Study (CCES)

- National & state polling data 2006-2020
- Harvard University, School of Government

Data Methodology

1. Random telephone polling
2. Volunteer pool to help balance underrepresented demographics
3. Weights to finish matching the sample to the adult population of California

* Note: Hispanics likely under-sampled, per study authors

Sample Sizes

Year	2016	2018	2020
	6,021	5,284	5,028

All respondents were likely voters in residing in California in the respective years.

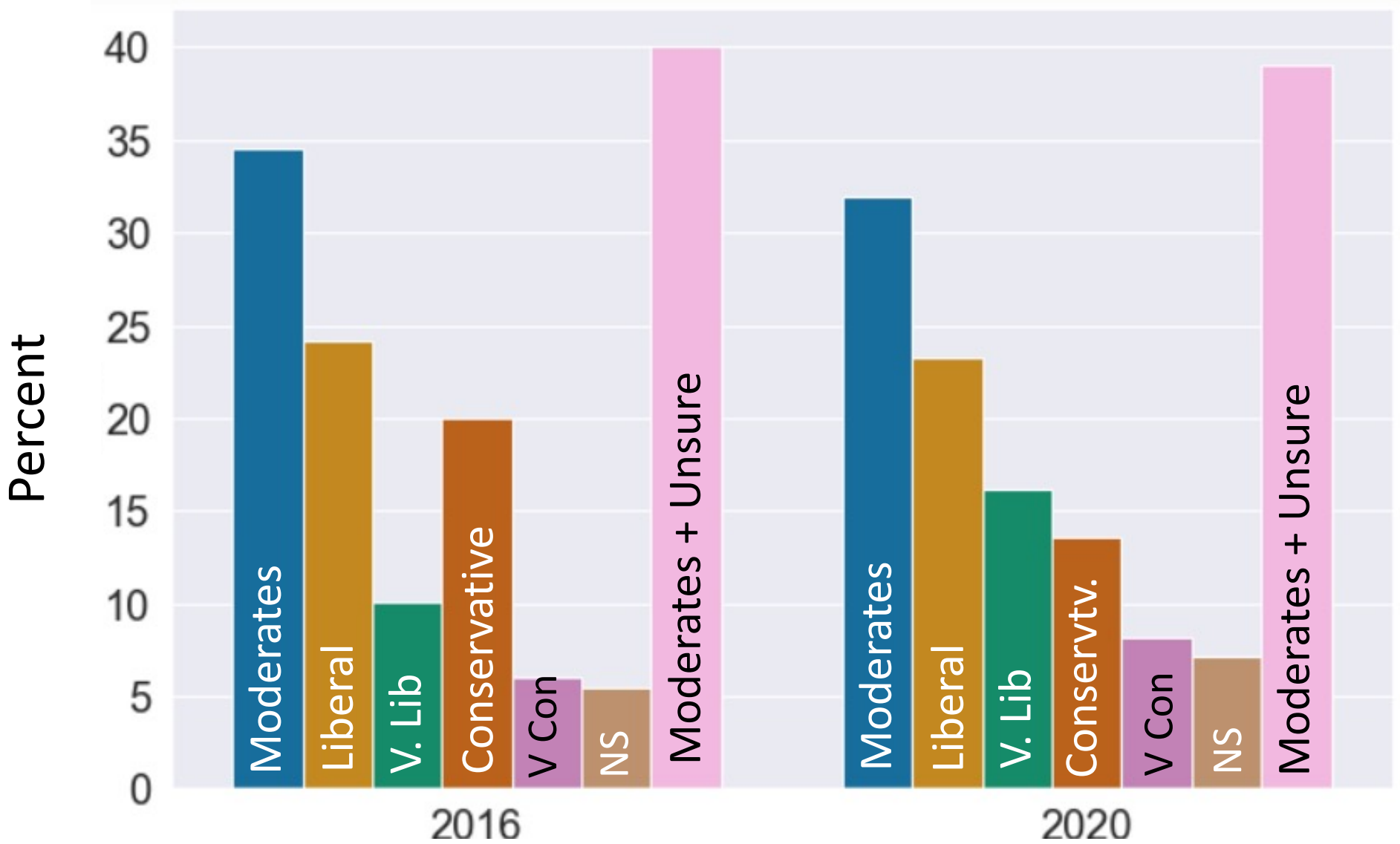
Variables

Var Name	Type	Name	Type
Age	Float	News interest	Likert
Approve Gov	Likert	Race	Categorical
Education	Categorical	Voted gov race	Categorical
Family income	Categorical	Voter Reg Status	Validated reg status
Gender	Categorical	Party Reg	Categorical
Ideology	Categorical		
Marital Status	Categorical		

EDA – Trends: 2016 - 2020

- Median age decreased from 47 to 45
- Percentage of whites decreased from 55.7% to 52.9%
- Percentage of married respondents decreased by 8%
- Percentage of never married respondents increased by 5%

Political Ideology 2016 - 2020



Our Methodology

All Classification models, since recall election was binary

- Logistic Regression
- Multinomial Naïve Bayes
- Random Forest
- Extra Trees
- Neural Network

Null Model

- Target: ”Who did you vote for?” – 2018
- Target Data: 62% Positive / 38% Negative
- Null Model: 62% Accuracy

Our Methodology

Best Models:

1. Voting Classifier
2. Random Forest
3. Neural Network

Voting Classifier

Ensemble Learning:

- Logistic Regressor
- KNN Classifier
- Multinomial Naïve Bayes
- We wanted our models to be interpretable

Voting Classifier

Metrics	KNN Class.	MN NB.	Logs Reg.	Vote Class.
Accuracy	81%	92%	92%	93%
Sensitivity	90%	94%	95%	96%
Specificity	65%	88%	88%	88%
Precision	81%	93%	93%	93%

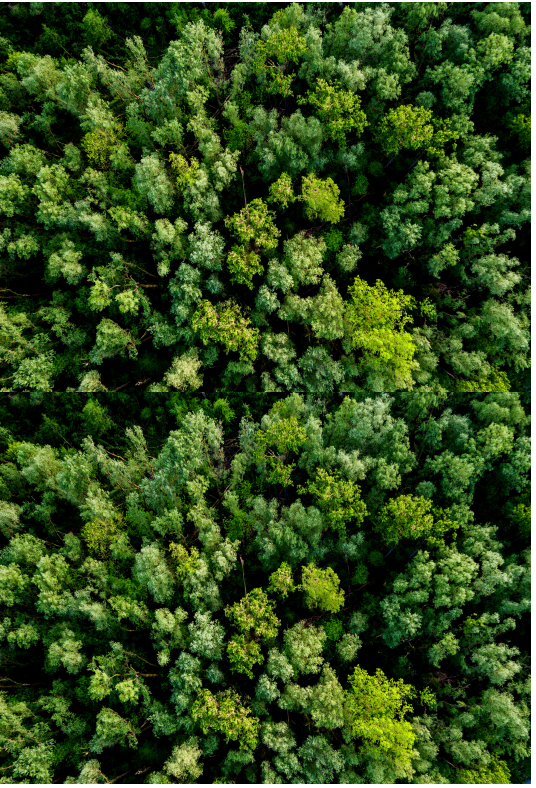
Voting Classifier

Coefficient	Percent Odds
Liberal	.943
Strongly App Gov	.854
Very Liberal	.760
Black	.197
Own Home?	.148

Coefficient	Percent Odds
Strongly Disapp Gov	-0.752
Pid3 repub	-0.673
Very Conservative	-0.320
Pid3 Independent	-0.250
Disapprove or Somewhat Dis Governor	-0.21

Voting Classifier

- Interpretation:
  - Misclassifications
  - Inference Information
  - Try Random Forest



Random Forest

Metric	Model 1 (%)	Model 2(%)
Accuracy – Train	100	93
Accuracy – Test	93.5	93
Specificity	88	89
Recall	97	96

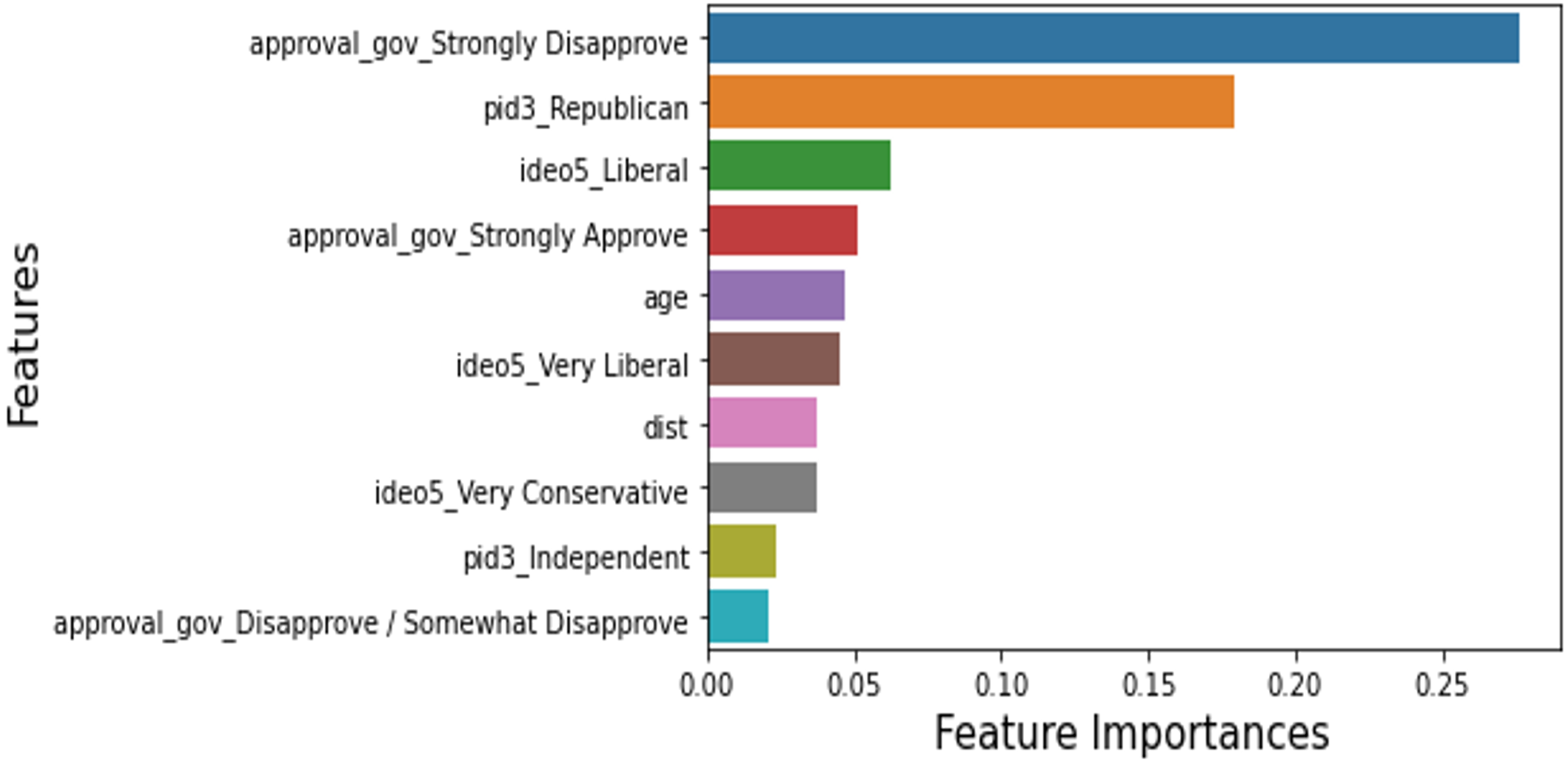
Neural Net

Metric	Early Stopping	Drop-out	ES & DO
Accuracy – Train	96	97.7	96
Accuracy – Test	93.6	92	92
Specificity	94	94.6	95
Recall	90	88	89

- 93% Accuracy
- 96% Recall
- 89% Specificity
- Bootstrap
- Max depth: 6
- Max features: 0.5
- Number of estimators: 100

Best Model: Random Forest

Random Forest: Top Ten Feature Importances



Model Test on Unseen Data

- Tested our model on Harvard **2020** data
- Average of predicted probabilities
  - Republicans - 31.3%
  - Democrats - 68.7%

Conclusions

- Model performed well on unseen data
- Similar models can be used to predict who someone will vote for
- Interpretability adds value to our modeling

Recommendations

- Improve accuracy using clustering techniques, and PCA
- Model in time series to see how changes to voter demographics change election outcomes

Problem Statement

Why do modern election results differ so much from the prognostications?

- Weighting
- Small sample sizes
- Sample interdependence
- Antiquated methods

Questions?