Predicting Error in Election Polling

Matthew Burgess

Introduction

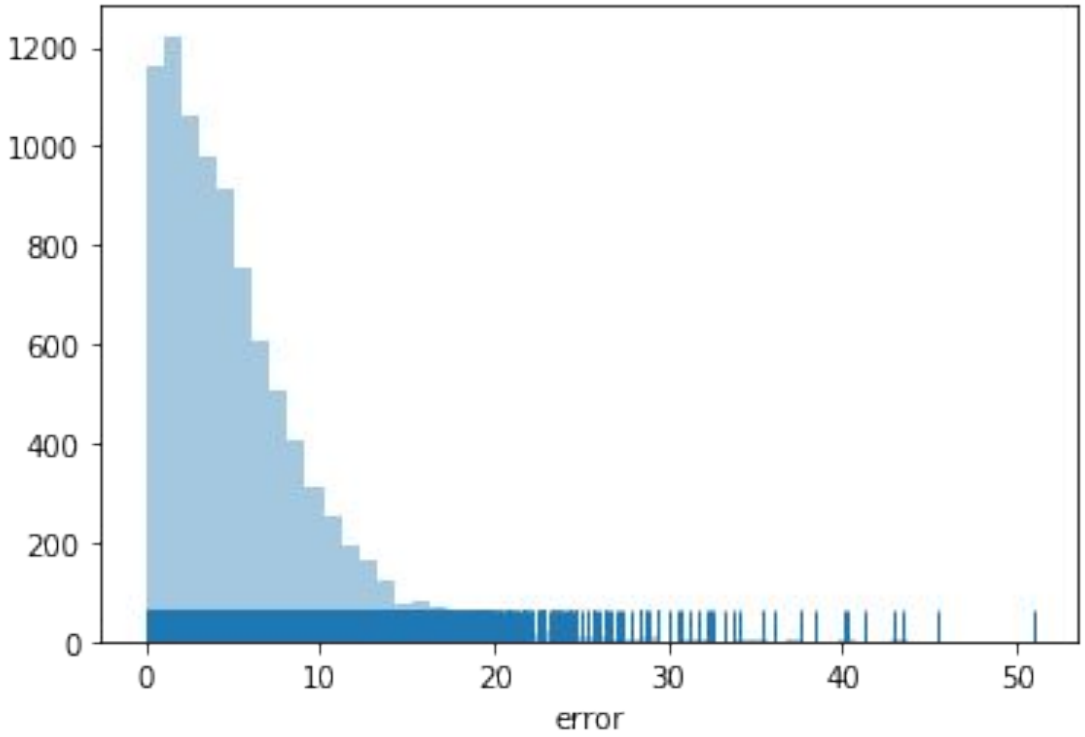
- Polls are useful for both campaigns and media.
- Having a method of measuring reliability provides clarity for campaigns and voters.
- The RELIABILITY of a poll is more important than the outcome.

Problem Statement: Create a regression model that predicts the error of a poll.

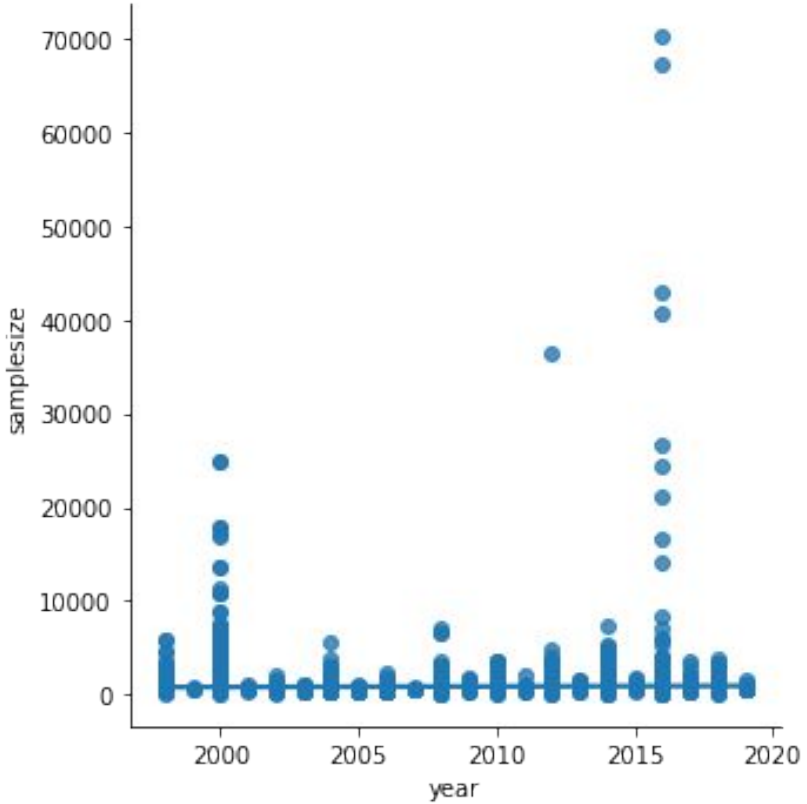
Data Collection

- Collected from FiveThirtyEight’s GitHub of data for their articles
- Approximately 9000 rows by 25 columns
- Ranges from 1998 to 2018

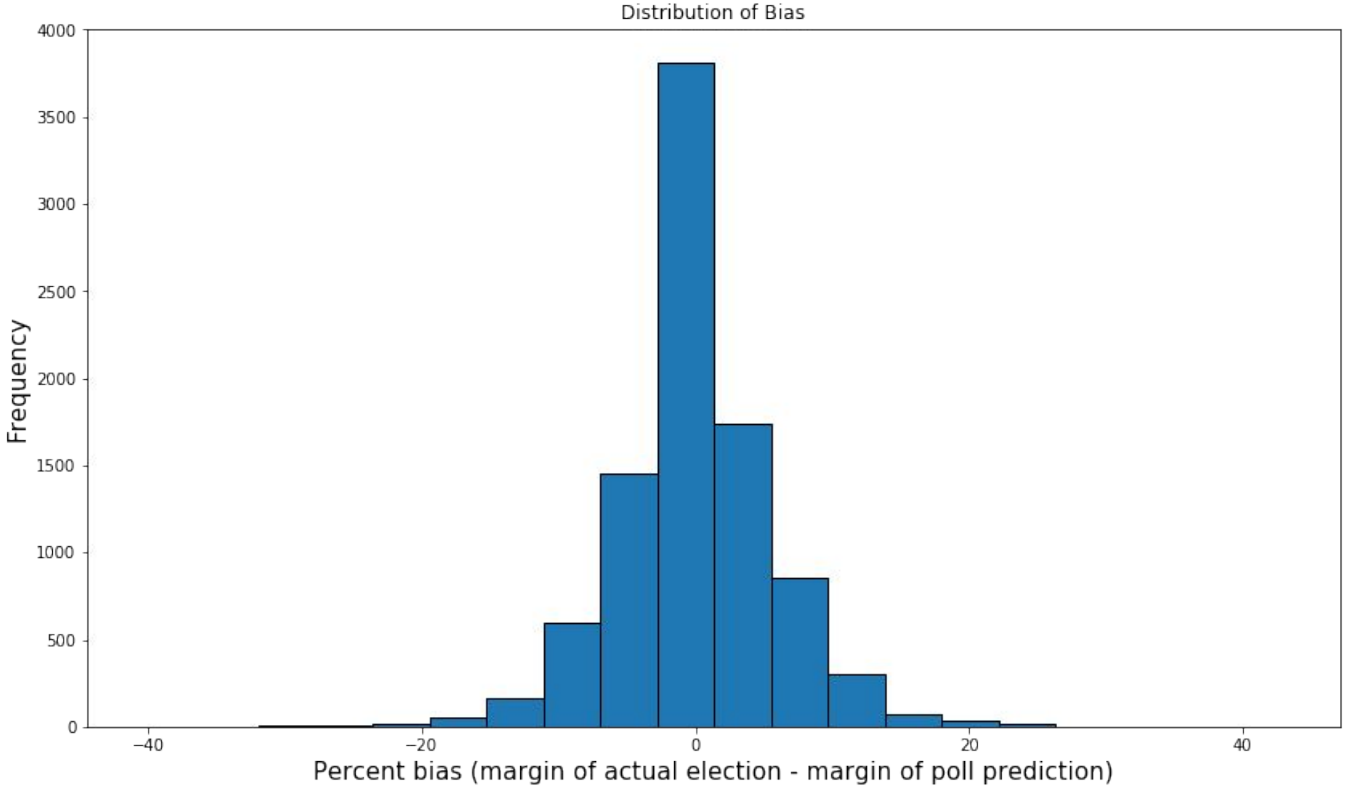
EDA: Distribution of Error



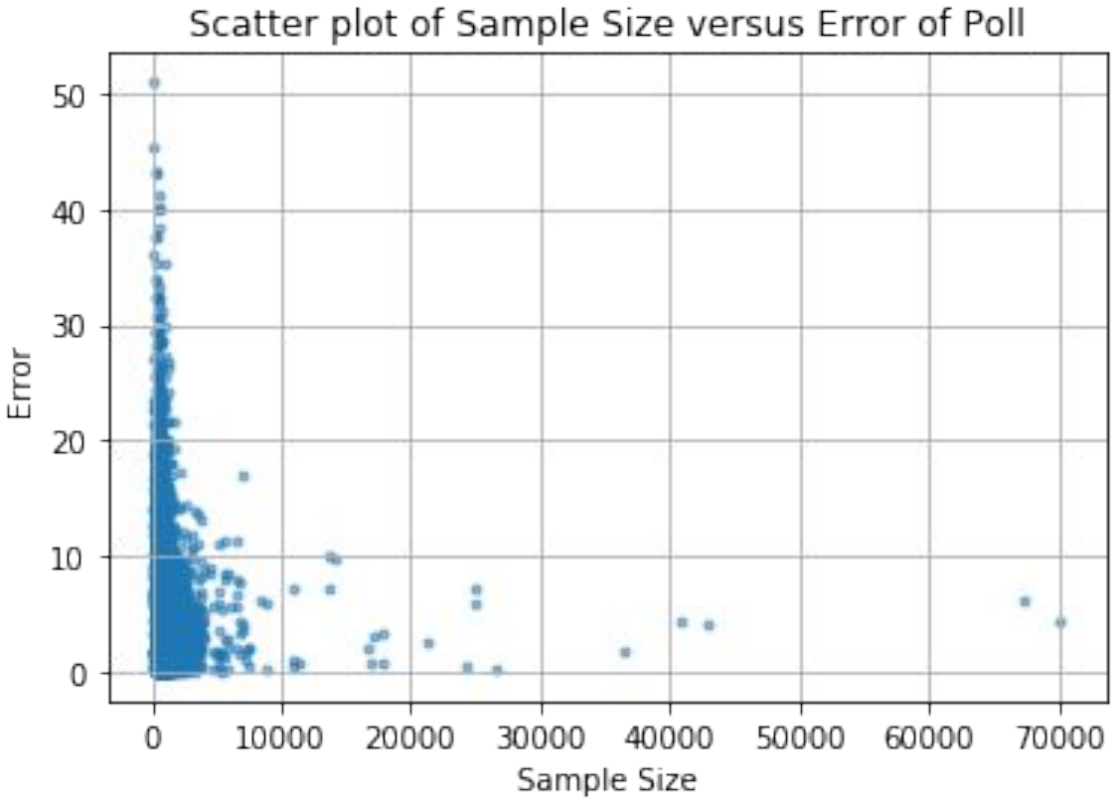
EDA: Distribution of Sample Size by Year



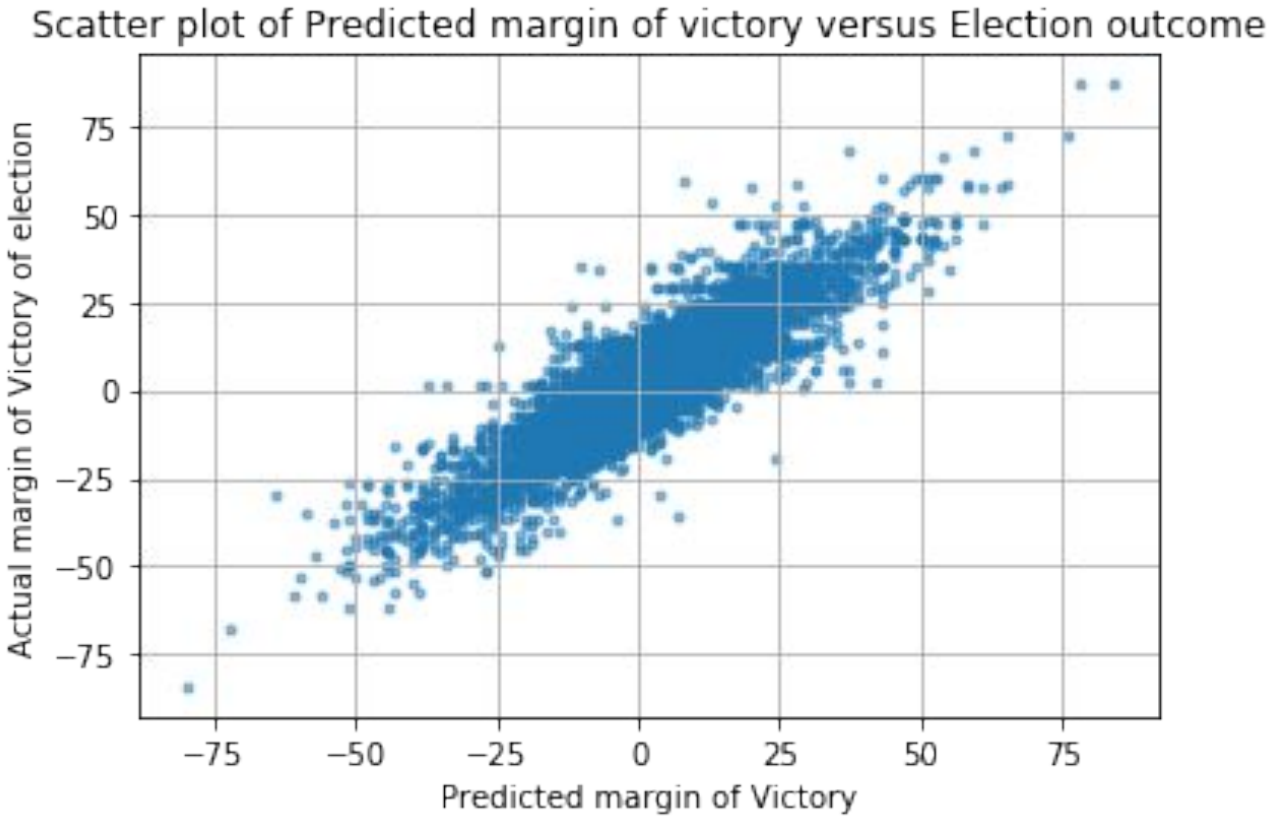
EDA: Distribution of Bias



EDA: Ruling out Collinearity



Visualizing Error as a Function: Predicted - Actual



Methods

- Linear Regression
- XGBoost
- Lasso Regression
- Ridge Regression

What is XGBoost? Why Use it?

- Extreme Gradient Boosting
- Decision Tree Method
- Minimizes model complexity through gradient descent
- Adds new models constantly
- Industry standard: computationally fast and effective

Results

Method	R-Squared Score	Mean Squared Error	Mean Absolute Error
Linear Regression	0.17545737688184448	18.73358721911903	3.1206837093578743
XGBoost	0.8274592710021751	3.920120931174652	0.9238756385858912
Lasso	0.07534753228877467	21.00808031698279	3.3709132867746283
Ridge	0.21706739446888657	17.78820868828203	3.0539818617526806

Conclusions

- XGBoost is most effective
- Has highest R-squared score of 0.82, and MAE of 0.92
- MAE of 0.92 indicates that we are also within 1% on our estimation
- None of the other models come CLOSE

Applications

- 85% of variability in polling error is captured by our model.
- Accurately predicting the error of a poll helps determine reliability of pollsters
- Ex: Running for Congress, and knowing if your paper’s polling can be relied upon to gauge status in the race. Can impact fundraising, endorsements, and momentum.