Springboard Data Science Intensive Course

Capstone Project 1

Modeling Polling Accuracy, from Governor to President

Marcus Bamberger

Table of Contents

Abstract

FiveThirtyEight's polling data was used to create several models to predict election outcomes and determine which characteristics of a poll are most predictive. The best models were a random forest classifier predicting if a poll would be within four percentage points of the final outcome and a logistic regression predicting maximum possible error of a poll. The most influential variables in these models were sample size and previous performance of the pollster.

1. Introduction

Tens of millions of Americans vote for their House and Senate representatives every two years and their president every four years. The last presidential election saw 128 million people cast their votes. Polling organizations like Gallup, the Pew Research Center, and YouGov attempt to poll America and determine in advance which way these elections will go. Due to the heterogenous nature of the country and cost of conducting wide, far-reaching polls, poll quality can vary greatly. This project seeks to examine a broad slice of polling data against election results to determine which polls are trustworthy and, more importantly, what makes a poll trustworthy.

Polling organizations conduct a wide variety of polls, from automated phone calls to online surveys to in-person interviews. These methods differ wildly in their costs to conduct and their accuracy. As phone scams and spam calls become more frequent, people answer the phone less, and the costs of polling increase. Responses to phone polls are down by around 70%. As more and more Americans find politics a source of interest and concern, it is necessary to know just what a poll means and how much it can be trusted.

This project seeks to forecast the quality of a poll by examining its methodology – how many people were surveyed, where the poll took place, how the pollster has performed in the past, et cetera. This is of interest to any political candidate who wishes to know how they are performing and how likely they are to win an election. The target audience is polling companies such as Gallup Poll and Harris Interactive, who would be interested in determining how to best improve their polling strategies with the least cost to themselves.

2. Finding Data

The website FiveThirtyEight collects and analyzes polling data from hundreds of different pollsters. The data tracks governor, House, Senate, and presidential races over the last 20 years. The data are publicly available via their GitHub page, located here. Their data are curated to remove falsified polls. Because FiveThirtyEight monitors pollsters and avoids selecting artificial data, we can avoid examining bad data in our models. The goal of this project is not to identify false polls, it is to determine which genuine polls are accurate.

FiveThirtyEight stores their data by election type, but each file is formatted in the same way. They track the date and year the poll was conducted, pollster identity, number of respondents, type of poll conducted, poll results, and candidates running. They also store pollster ratings, a letter grade indicating the overall quality of the pollster, but this data was discarded. This project is working from scratch, not evaluating FiveThirtyEight's metrics.

The data were compared against final election results and election dates, taken from the New York Times and Wikipedia. Candidate party affiliation was also taken from the New York Times and Wikipedia where necessary. FiveThirtyEight's dataset does sometimes clarify party affiliation, but their 'cand_name' column can contain names, such as 'Kerry', names followed by party affiliation, such as 'Lamont (D)' or only party affiliation, such as 'Republican'. In order to track polling error by party or partisan race, we found and added candidate affiliations to the dataset.

The 2018 data were formatted differently than the rest of the data: they tracked one candidate per poll per row, instead of the prior convention of one poll per row. The 2018 data were cleaned to match and the data were merged. Several variables of interest, such as time between poll and election, or state the poll was conducted in, were also added to the dataset. The Jupyter Notebooks containing the code used to reformat and clean the data can be found here.

The other issue that arose when tracking and wrangling the data were the intermittent presence of single-party races. Because we suspected that partisanship would be a valuable indicator of a poll's accuracy, we added a new column, 'partisan_race', to indicate if the race

was between a Republican and a Democrat or not. Races between three candidates with at least one Republican and at least one Democrat were also classified as partisan. Despite running as an independent, Bernie Sanders is classified as a Democrat here given his strong ties to the Democratic party.

The primary dataset was cleaned and ready to be analyzed: Over 10,000 polls taken over 20 years of polling, tracking such factors as year, sample size, methodology, error and bias in the poll, and the state the poll was taken in.

3. Statistical Data Analysis

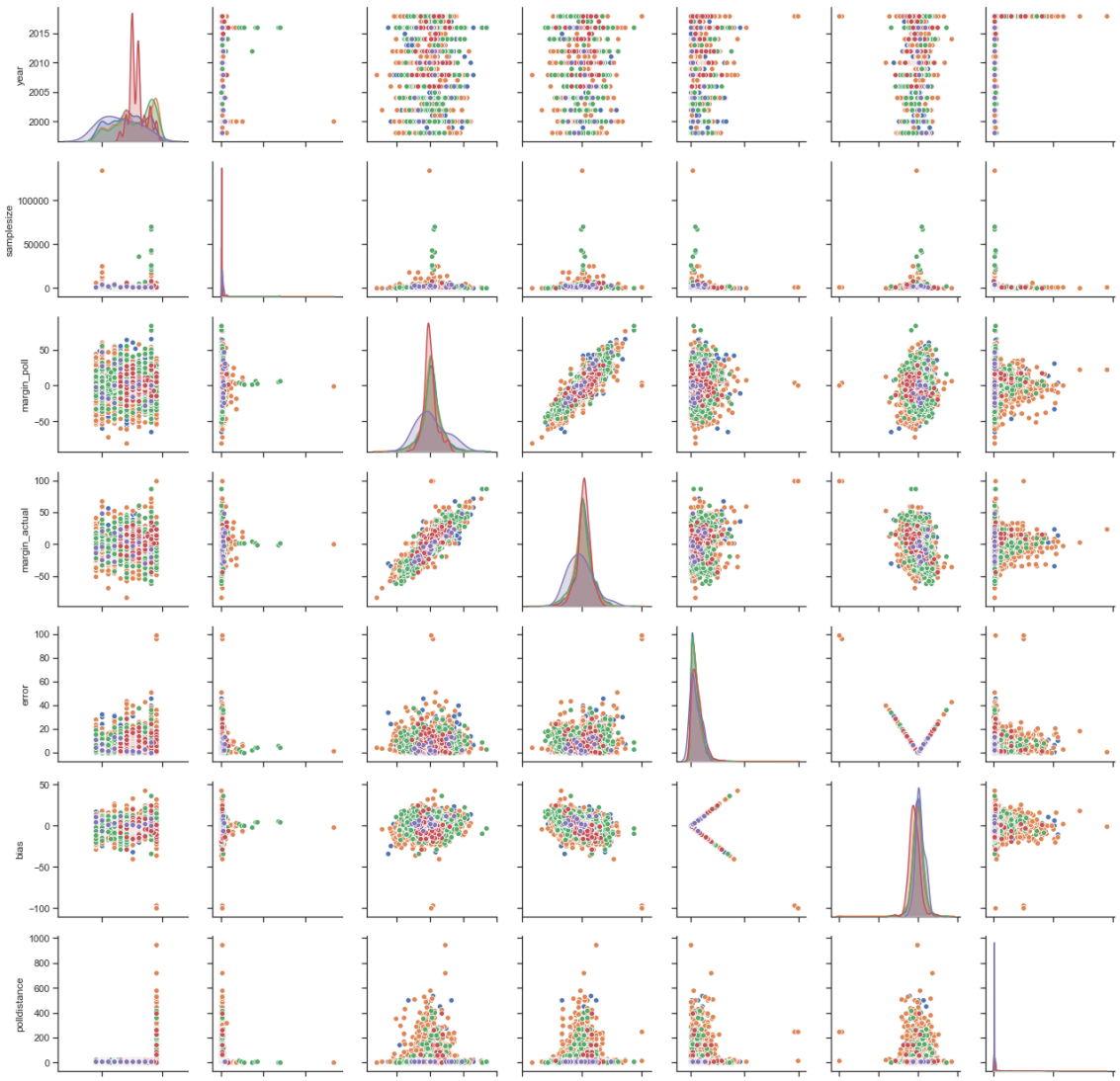

Figure One: Pairplot of the most relevant statistics in the capstone dataset. Points are colored by type of poll.

The goal of the initial examination was to find the variable that exerted the most influence on the error and bias of a poll. Bias is calculated with the following formula:

$$Bias = \left(Candidate\ 1\ \%_{poll} - \ Candidate\ 2\ \%_{poll}\right) - \left(Candidate\ 1\ \%_{true} - Candidate\ 2\ \%_{true}\right)$$

where % indicates the percentage of the total popular vote they received. The sign of the bias indicates which candidate did better than expected. Per FiveThirtyEight convention, in partisan races a positive bias indicates the Democratic candidate outperformed their polls, and a negative

bias indicates the Republican candidate outperformed their polls. Error is the absolute value of bias, and indicates how accurate the poll was without examining partisanship.

Three of the first tested hypotheses tested were that sample size correlates negatively with error, polling data improves as the election approaches, and the type of poll strongly impacts the error of the poll. None proved to be exactly correct. The Jupyter Notebook code to construct these graphs can be found here.

a. Sample Size

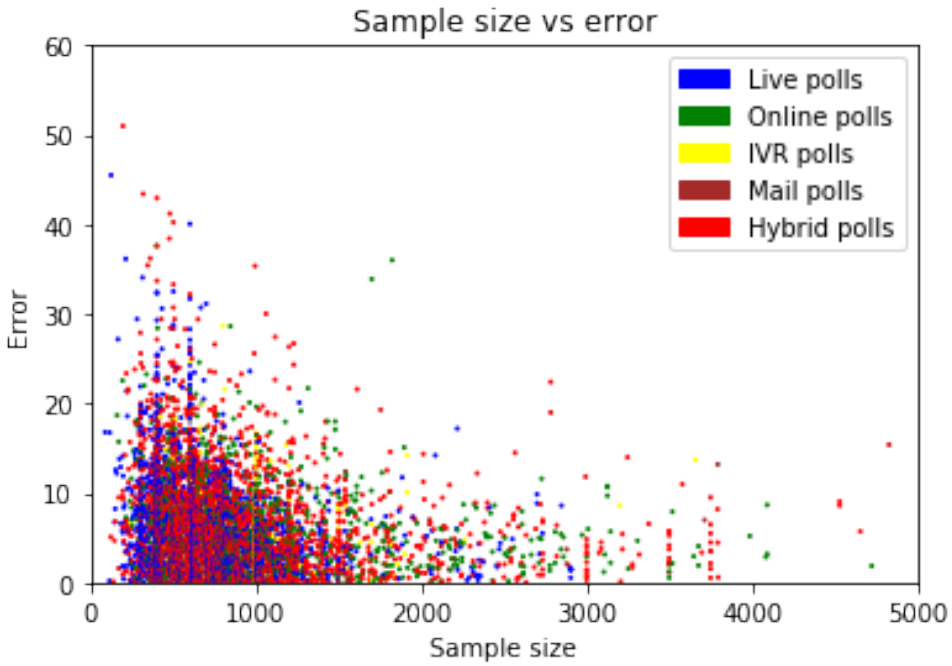

Figure Two: Scatter plot of sample size versus polling error, colored by poll type.

As this poll indicates, even polls with tiny samples can have extremely low error. It would be more accurate to say that sample size reduces maximum error: It is very difficult to be badly wrong with 2000 respondents, but you can be very lucky with 10 respondents. This poll also tracks polling type, but there is no clear correlation of poll type with error here. We'll dig deeper into the effect that poll type has on error later.

It's worth taking a moment now to explain how a poll can plausibly predict anything at all. The average sample size here is 850 people, and America contains around 250 million people of voting age, per the United States Census. In short, demographics tend to vote together. If

seven in ten people over 65 say that they'll vote for candidate A over candidate B, it's a reasonable bet that around 70,000 out of 100,000 people over 65 will, too. It's also important to weight your forecast by voter count: If you expect that twice as many 65+ people will vote as 18-25 people, but your poll saw equal responses from both demographics, you'll want to weight the 65+ responses twice as heavily. We'll discuss this more later, with regards to weighting and polling cost.

b. Poll Distance

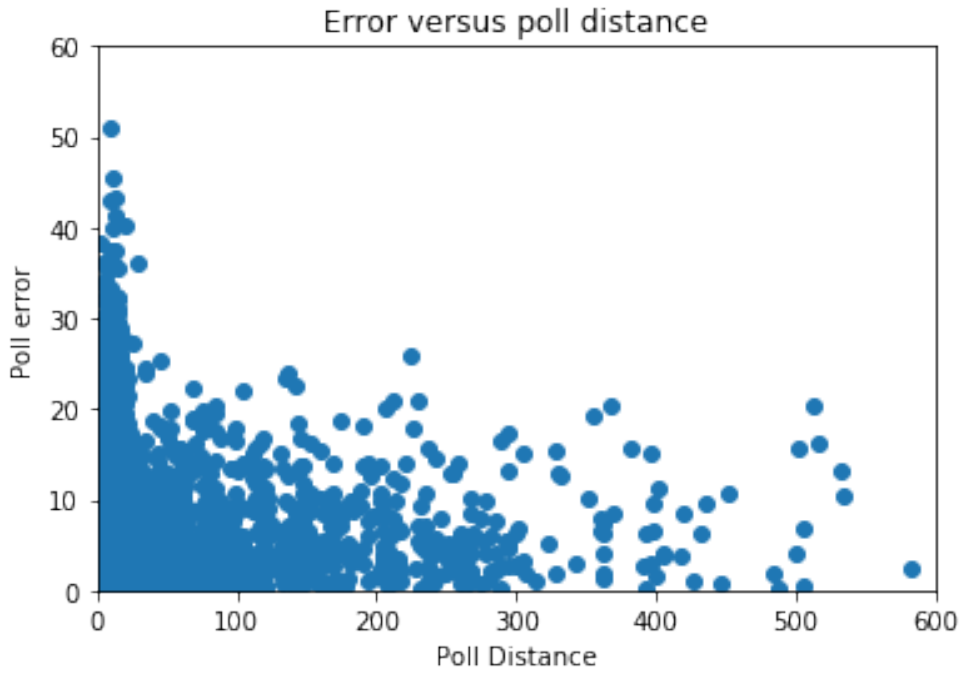

Figure Three: Scatter plot of error versus poll distance.

We expected that polls taken closer to the election would be more accurate, as there would be less time for major scandals, debates, and other events to shift opinion. Instead, we observed a massive spike in inaccurate polls immediately before the election. Initially, we theorized that this was due to 'Herding', the practice of adjusting polls taken late in the polling cycle to more closely reflect prior polls. This is typically done by adjusting the weighting system until results match expected results. FiveThirtyEight also tracks herding in their database, but again, we chose to work from scratch for this project. Herding would hold accuracy at a fixed level, but not reduce it, so it does not properly explain these observed results.

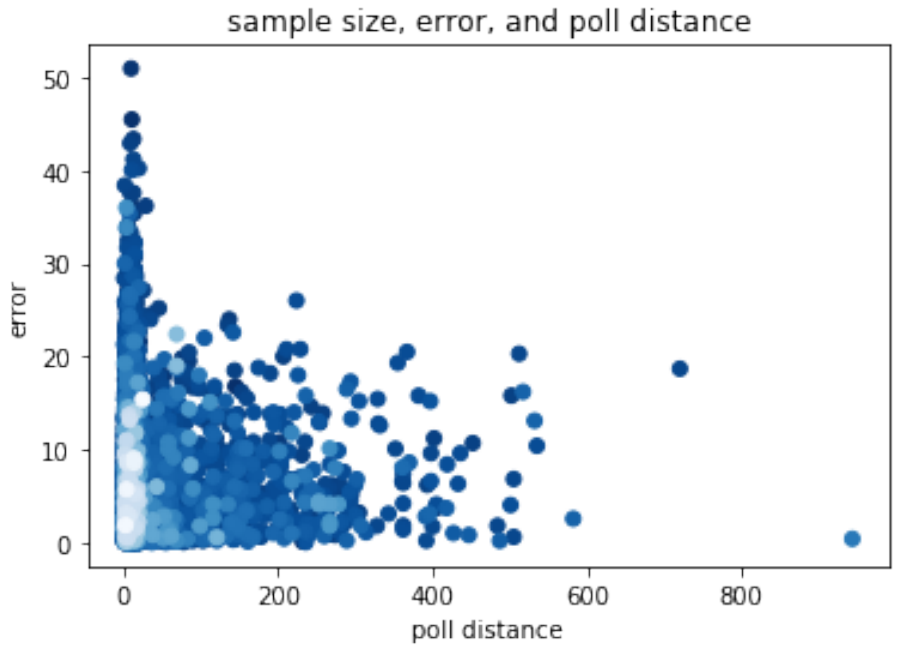

Figure Four: Scatter plot of poll distance against polling error, colored by sample size.

A second analysis, tracking sample size against poll distance and error, provided a better explanation. Many of these last-minute polls have low sample sizes. It is not the case that polls taken at greater distance to the election are better, but that sample size is a confounding variable here.

c. Poll Type

We theorized that the type of poll conducted would strongly influence the quality of the results. Mail polling and IVR (Interactive Voice Response – robocalls) polls would be expected to reach different demographics than online polling, for example. Costs of polling also vary greatly by type of poll: if there were no benefits to paying humans to call people instead of leaving it to robots, why would anyone do it?

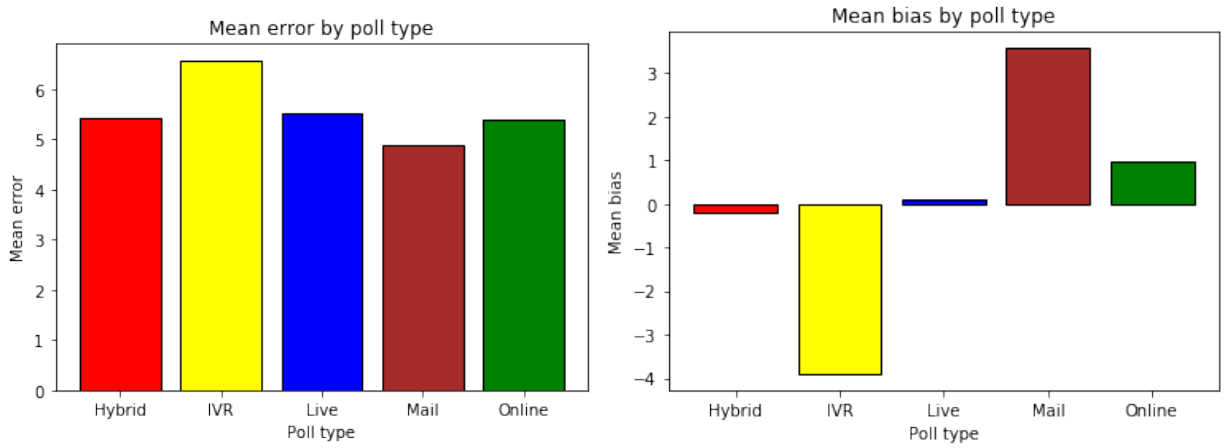

Figure Five: Mean error and bias by poll type.

We see that poll type strongly predicts bias, but not error: All polling methodologies are likely to miss their mark by the same margin, but some polls tend to overrate Democratic candidates, some overrate Republican candidates, and some show no consistent bias. Poll type is therefore of minimal use when evaluating absolute poll error, but significantly more valuable when predicting the bias of a poll.

3.4 Location

The state a poll is conductive in is strongly correlated with the bias of the poll.

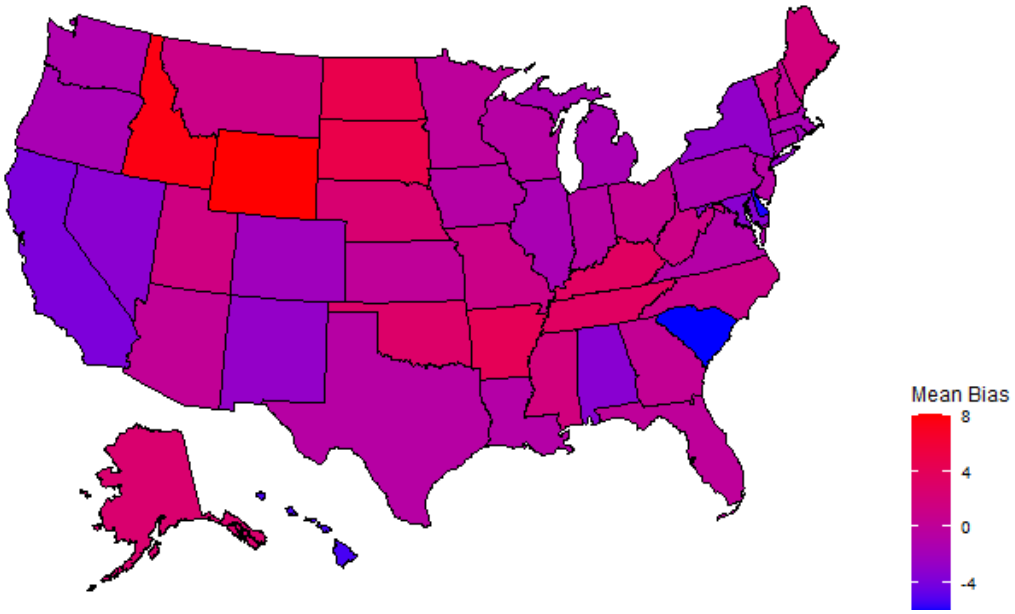

Figure Six: Average bias by state.

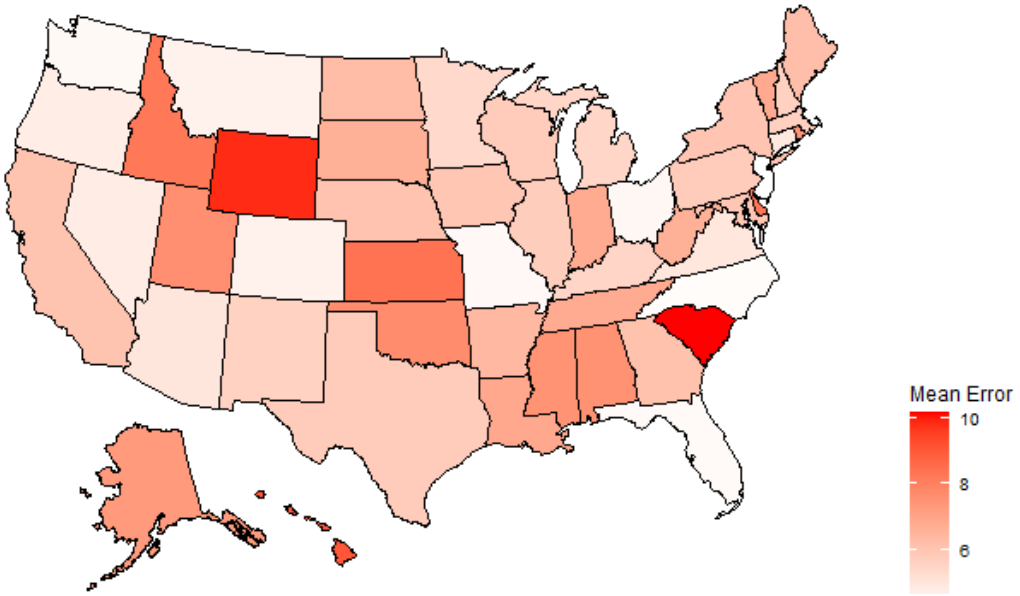

Figure Seven: Average error by state.

We can see that polling varies significantly by state: South Carolina and Wyoming are both among the least accurate, but Wyoming tends to underrate the Democratic candidates and South Carolina tends to overrate them. Additionally, the maximum Republican bias is greater than the maximum Democratic bias by ~2.5 percentage points. Tracking all 50 states in a model would introduce significant dimensionality into the results, but a simplified system that divided states into 2-4 groups based on their mean error has the potential to improve results without overcomplicating the model,

3.5. Prior Performance

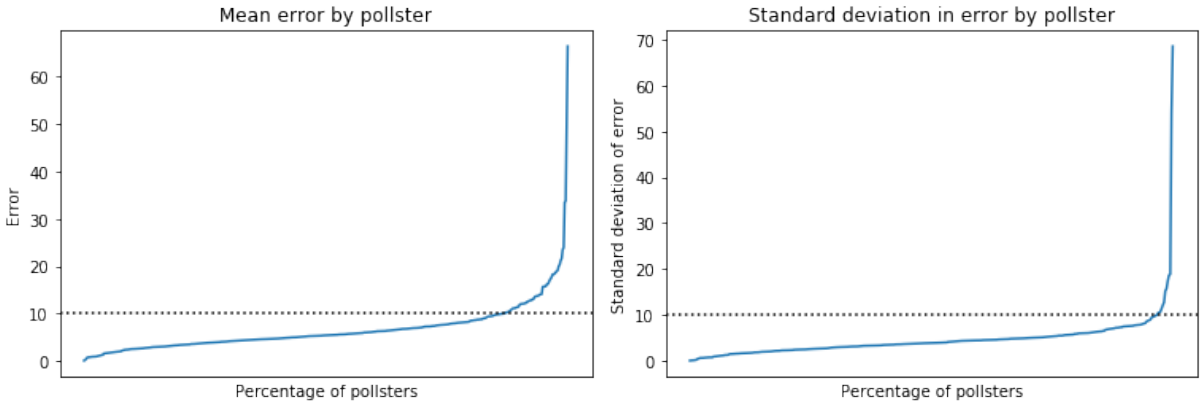

Figure Eight: Mean error and standard deviation of error per pollster.

Often, a strong indicator of future performance is past performance. Of the 390 polling organizations with more than one poll recorded in FiveThirtyEight's dataset, 377 of them, or 96%, had a standard deviation of polling error less than 10 percentage points. In other words, pollsters tend to have similar levels of accuracy across multiple polls. We can track prior pollster performance to better predict their future accuracy.

This is only possible for pollsters with more than one recorded poll, of course. While we can still make assumptions about future polls based on the error of their one recorded poll, we cannot compute a standard deviation from a single point of data. This is a minor issue, given that only around 95 polls in the dataset were contributed by single-poll pollsters – less than 1% of the total. Still, it is worth noting that our models will be less accurate when examining polls from new pollsters.

Other variables that were examined and found to have no significant effect on the accuracy and precision of the poll were the predicted margin of victory – if the race was thought to be close or not – and the type of race. Governor's races and Senate races are equally accurate, and it's just as hard to precisely call the outcome in a close race as it is in a landslide.

4. Modeling

4.1. Ordinary Least Squares

After examining the data, the clearest single-variable correlation with error appeared to be sample size, which correlated negatively with the maximum error of a poll. Determining the relationship between error/bias and polling sample size was complicated by the fact that sampling size described a maximum error, but did not restrict the minimum. In other words, it was possible to sample 50 people and make an error-free poll, but a poll with 7500 respondents was very unlikely to be badly wrong. Modeling error against unadjusted sample size did not produce satisfactory results. It was necessary to further clean and prepare the data for modeling.

We binned the data by sample size in 50-person intervals and took the maximum error of each bin.

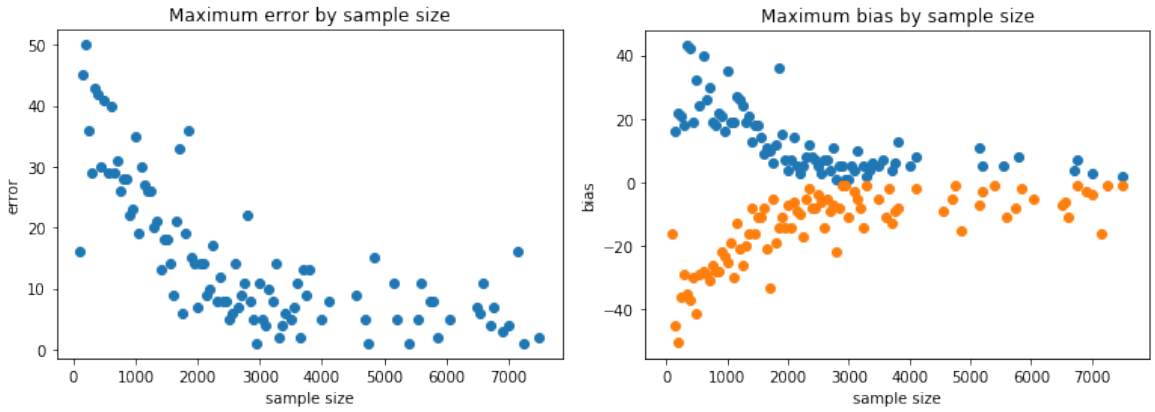

Figure Nine: Binned error and bias against sample size.

We can see that maximum bias and error have a roughly linear relationship with sample size up to around 4000 respondents. Increasing sample size beyond this threshold does not significantly improve accuracy.

After examining several models, the statsmodels OLS was found to be a good fit for the data. A simple linear regression of sample size up to 4000 respondents against error yielded an adjusted R-squared of 0.655 and F-statistic of 142. With a p-value of zero, it was clear that these factors were correlated. It was unclear if the relationship would be better modeled with linear or logistic regression, so both were tested. To improve the model further, we experimented with incorporating the prior performance of the pollster that performed the poll.

Independent Variables	Dependent Variable	R-squared	Adjusted R-squared	F-statistic
Sample size	Error	0.660	0.655	141.8
Sample size	log(Error)	0.610	0.604	114.0
Sample size, pollster_std	Error	0.836	0.832	184.0
Sample size, pollster_std	log(Error)	0.863	0.859	226.2
Sample size, pollster_error	Error	0.722	0.714	93.5

Independent Variables	Dependent Variable	R-squared	Adjusted R-squared	F-statistic
Sample size, pollster_error	log(Error)	0.802	0.797	145.9
Sample size, pollster_std, pollster_error	Error	0.842	0.835	125.8
Sample size, pollster_std, pollster_error	log(Error)	0.867	0.862	154.7

Table One: Testing OLS modeling

After some testing, the best OLS model was found to be sample size and standard deviation of pollster error against log of error. Adding the mean error of the pollsters to the model marginally improved the R-squared but greatly reduced the F-statistic, and was overall unnecessary.

Taking the log of the error initially appeared to reduce the R-squared and F-statistic values, indicating that the error was linear and not logarithmic, but once prior performance was factored into the model, the logarithm of the error performed better. We can conclude that error more closely follows a logarithmic relationship with prior performance and total respondents.

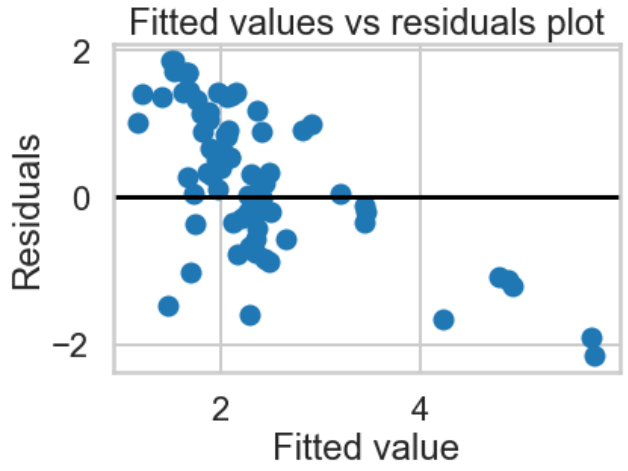

Figure Ten: Residuals plot of final OLS model

As this residuals plot indicates, the residuals are not randomly distributed: We have not eliminated all bias. However, the lack of a clear curve or other pattern to the residual plot shows that this model lacks major bias.

This analysis was done with an arbitrarily chosen binning of 50. After determining that there was strong correlation between sample size, past pollster performance, and error, the model was re-tested for a range of bins up to 100 in size.

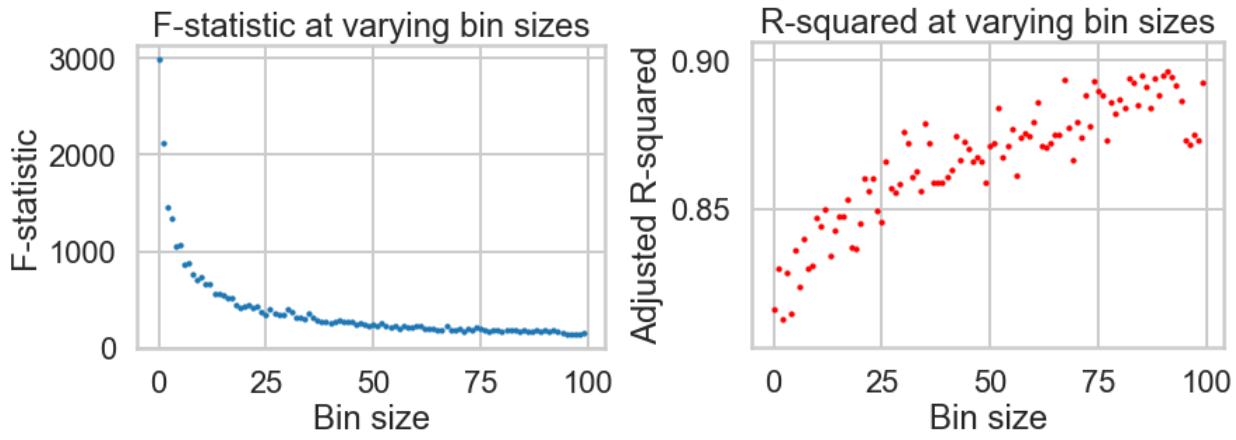

Figure Eleven: R-squared and F-statistics for varying binnings

Running the model with smaller bin sizes reduces the R-squared value but increases the F-statistic. The gains in F-statistic are significantly larger than the losses in R-squared, as shown here:

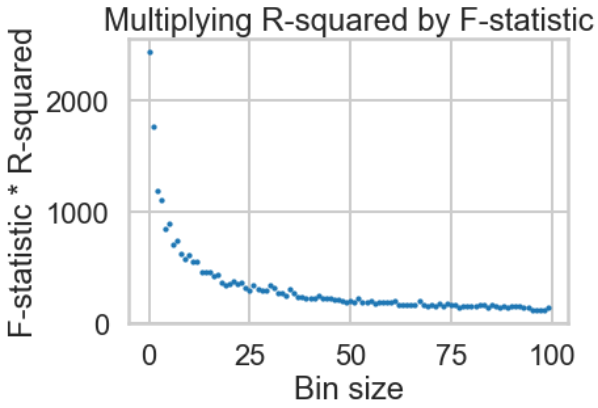

Figure Twelve: Product of R-squared and F-statistic for varying binnings

Although large binnings increase the R-squared of the model, they damage the F-statistic and risk oversimplifying the model. At a binning of 100, only 40 data points are present. If we set a minimum R-squared of .85, we can still achieve a F-statistic of 656 and bins of size 13. This model predicts 85% of the variance in maximum error with sample size and pollster standard deviation.

This OLS model does well for predicting maximum error, but fares significantly less well for predicting exact error of a given poll. When applied to the dataset of all polls with no binning

for error maximization, R-squared fell to 0.516. Additional, more robust modeling is needed to accurately predict the actual error of a poll, not just the maximum error.

The Jupyter notebook containing the code used to build this model can be found here.

4.2. Naïve Bayes

Modeling maximum error does have some value: With a simple OLS model we can predict the minimum accuracy of future polls from a given pollster. However, it would be more valuable to predict the margin of error of a given poll. With more complex modeling, this is an option. For datasets of this size and complexity, the best options are typically classification models, which split data into two distinct groups based on their characteristics. 46.6% of all polls in the dataset had less than 4 percentage points of error: For an election where n% of the vote went to candidate A, they predicted A would receive between n-4 and n+4 % of the vote. We examined classification boundaries such as less than/greater than 4 percentage points of error, as well as other splits such as 2%, 1%, and log(2%). Finally, we tested classifying polls as being correct or not: if they gave a higher vote margin to the candidate who would eventually claim victory.

We began modeling exact error with a Naïve Bayes model. Although it is not the most robust classification model overall, it was a strong choice for initial data exploration as we determined which variables and variable permutations would best correlate to error or log(error) of the poll. The major issue we encountered when modeling Naïve Bayes was that the model tended to aggressively classify all polls as belonging to the larger of the two classifications. This prevented us from constructing high-quality Naïve Bayes models for polls with more or less than 1% error, as those polls composed only 12% of the dataset, polls with 2% error, which composed 25% of the dataset, or polls that incorrectly predicted the result, as those composed ~20% of the dataset. These problems were not immediately apparent because the R-squared of the models did not sufficiently penalize this.

For this reason, we began scoring our models not by R-squared or F-statistic, but by F1 score, which measures the accuracy of a test in terms of its ratio of true and false positives and negatives. The F1 score is computed as:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

where Precision is the ratio of true positives to total positive results and Recall is the ratio of true positives to total positive items.

Independent Variables	Dependent Variable	Training accuracy	Test accuracy	F1-score
Samplesize, pollster_std	4% error	0.545	0.558	0.386
Samplesize, pollster_error	4% error	0.565	0.579	0.449
Samplesize, pollster_std, pollster_error	4% error	0.560	0.572	0.450
Samplesize, pollster_std	1.25% log(Error)	0.583	0.583	0.231
Samplesize, pollster_error	1.25% log(Error)	0.603	0.569	0.300
Samplesize, pollster_std, pollster_error	1.25% log(Error)	0.591	0.601	0.377
Samplesize, pollster_std, pollster_error, state	4% error	0.537	0.537	0.422
Samplesize, pollster_std, pollster_error, poll type	4% error	0.566	0.574	0.445

Table Two: Scores of varying naïve Bayes classification models.

The best naïve Bayes model used sample size and pollster mean error to predict if a poll came within four percentage points of the correct result. Adding pollster standard deviation or pollster standard deviation and methodology did not meaningfully impact f1 score and slightly reduced test accuracy. We can conclude that these variables are closely tied to the final error of the poll. The actual predictive power of the naïve Bayes was too low to be useful, however: At .450, the model is slightly worse than random. A more robust model is necessary to handle these predictions.

4.3. Categorical Random Forests

After finishing work with naïve Bayes models, we turned our attention to random forest classification modeling. This model proved to be more robust at handling multiple independent variables, and we were able to construct decision trees incorporating several additional variables besides sample size and pollster error.

The three main factors that determine the results of a random forest model are the number of trees, which is limited only by time and computation power, the depth of the trees, and the decision factor. For initial exploratory modeling, the number of trees was set to 1000, and for final modeling, 10000.

The default criterion for node splitting in a random forest model is gini impurity, and there was no reason to alter this: the entropy criterion performed worse. As such, the random forest models we worked with here were refined by varying their independent variables, dependent variables, and the depth of the trees. This can be done directly, by setting a maximum allowable depth, indirectly, i.e. by setting a minimum node size, or with a combination of both.

Because random forest models are complex, they are prone to overfitting their data. Setting a maximum depth restricts their complexity and prevents them from overfitting. We can measure this overfit by taking the difference of the training and test data sets.

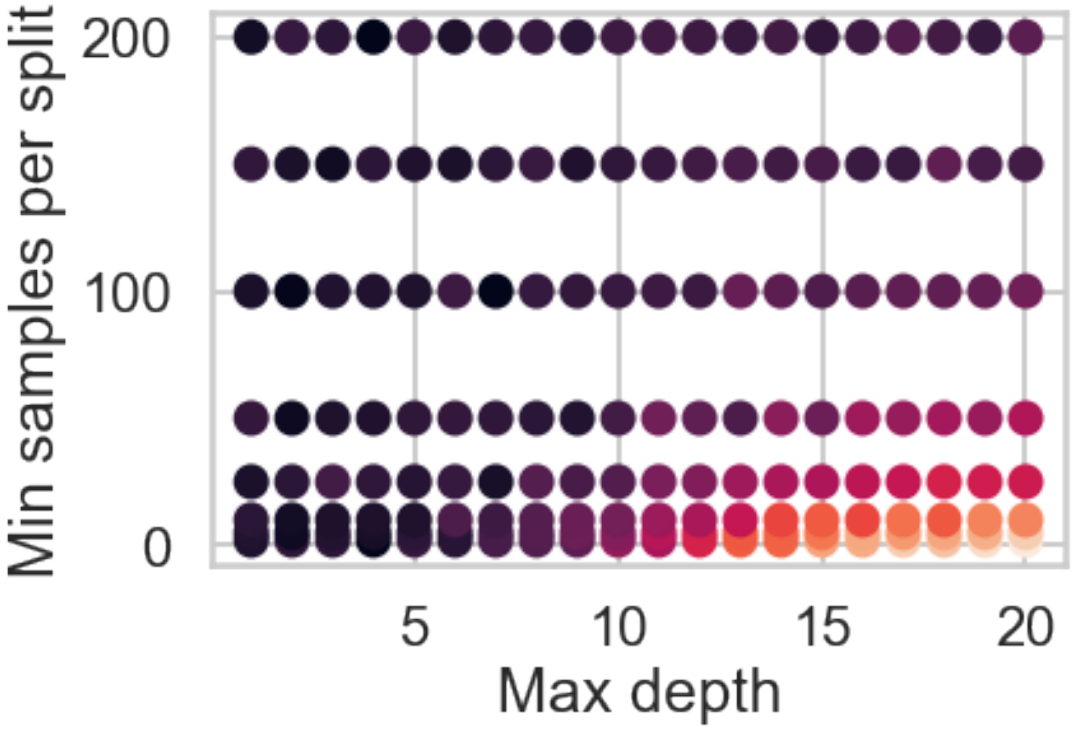

Figure Thirteen: Tracking overfit of 140 random forest models with different parameters.

Restricting minimum samples per node to 100 removes almost all overfit, even when we allow the trees to reach extreme depths, but this comes at significant cost to the overall accuracy of the model. Allowing the trees to split more freely dramatically increases overfit at higher depths, but there is no time for the models to develop overfit when restricted to shorter trees. We can see the accuracies of the different models here.

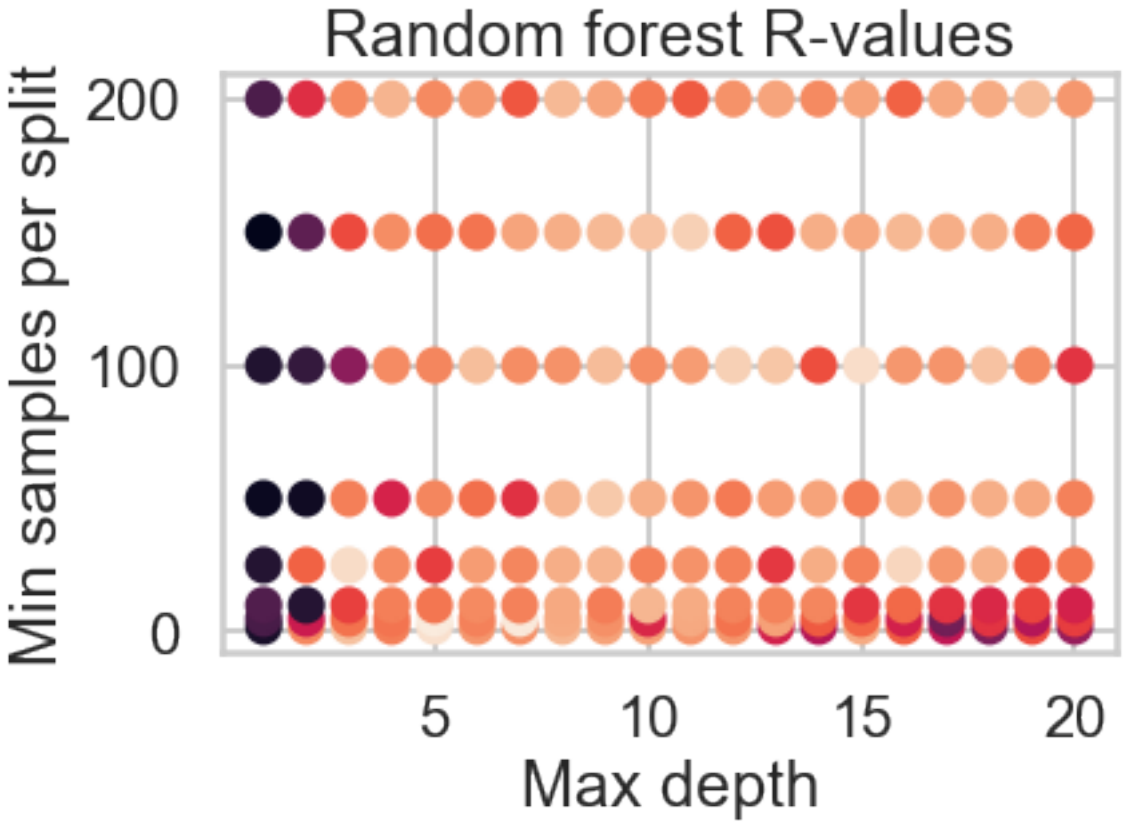

Figure Fourteen: Tracking R-values of random forest models with varying parameters.

The accuracy of the random forest model is not significantly increased by raising the maximum depth of the trees above 5 for any node size restriction. In fact, raising the maximum depth too high with no restrictions on node size produces a worse model than one with those restrictions in place. We can also see that restricting sample size has little effect on the adjusted r-squared value.

For many variables, such as the state the poll was taken in or the time before election, passing them into the random forest model unedited produced poor results. The variables introduced too much dimensionality into the model for too little result. However, we were able to extract value from these models by reducing them to a few categories, sometimes as little as two.

This reduced dimensionality conveyed the most important information from the variable without going into unnecessary detail.

For example, a random forest model that tracks poll sample size, time before election, and standard deviation of pollster error raises its f1 score by .016 when the state variable is added, and by .076 when the state variable is added as a trinary variable: five times higher.

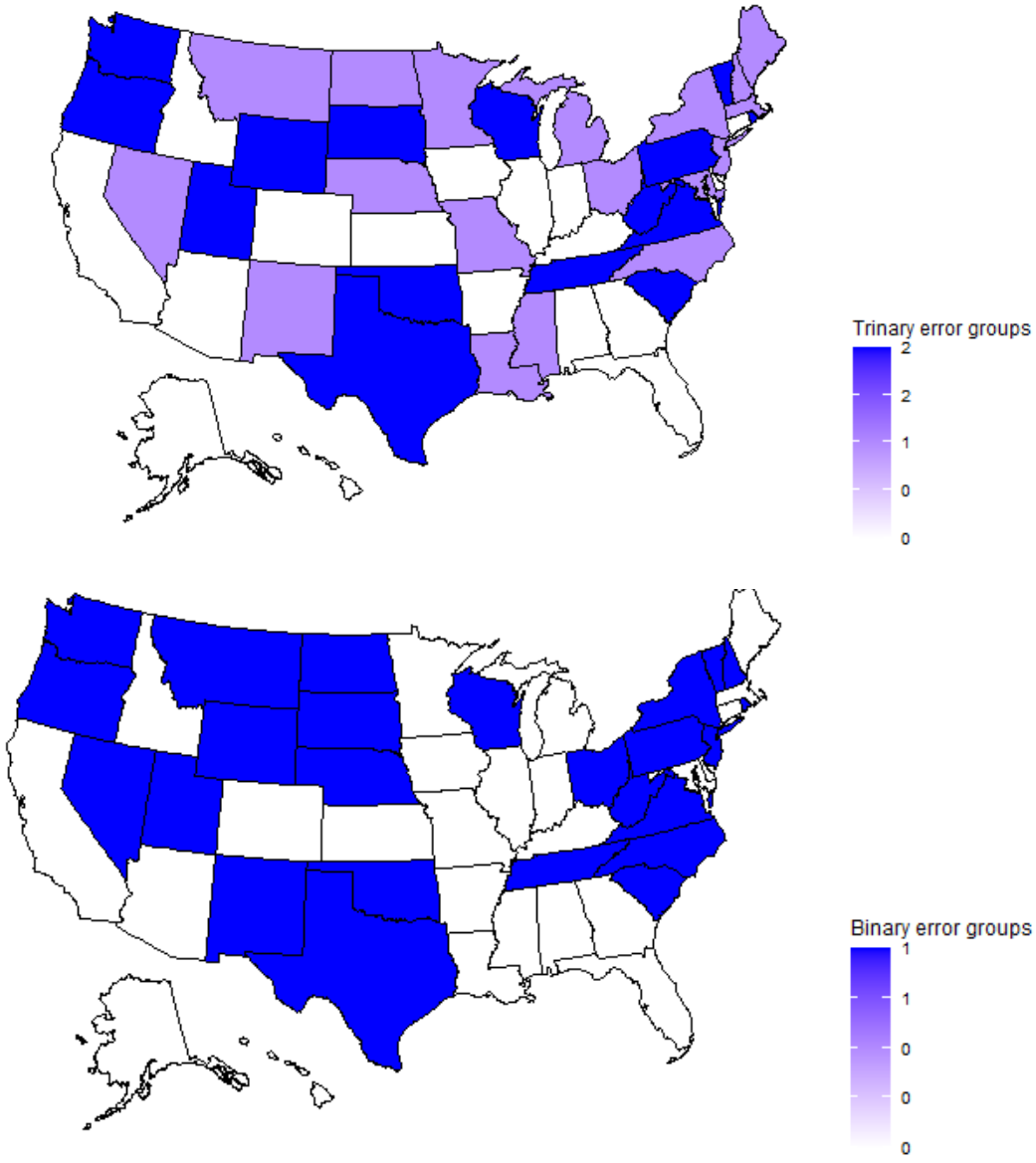

Figure Fifteen: Reducing the state variable to more manageable dimensionality.

Tracking R and f1 scores for varying dimensionality of the state variable showed that the best way to incorporate state data into the random forest model is as a binary classification: Either the state is in the top 50% in mean error, or it is not. Something similar is observed with the continuous variable of time before election: Training the model on a binary split between the 25

days prior to the election, when poor polls proliferate, and all other times, produces a better model than if the continuous distance variable is passed in. However, ultimately the best model did not include poll distance.

We also experimented with a partial sampling of the data for the 'right_call' classifier: approximately 20% of the dataset contained polls that did not correctly forecast final election results, so we selected an equal number of correct polls at random and trained the model on that data. Each model trained this way only received around 40% of the total dataset for training, but they did not aggressively classify the data. These models were an improvement over the previous models, but suffered from a lack of data. Training the random forest model to classify based on 4% error performed better, as this feature split the entire database evenly.

Testing different random forest models yielded the following results:

Independent Variables	Maximum depth	Minimum node size	Training accuracy	Test accuracy	F1-score
Samplesize, pollster_error	5	10	0.596	0.584	0.454
Samplesize, pollster_error	5	25	0.604	0.577	0.463
Samplesize, pollster_error	12	10	0.713	0.581	0.542
Samplesize, pollster_error	12	25	0.683	0.594	0.590
Samplesize, pollster_error, pollster_std	5	10	0.597	0.591	0.590
Samplesize, pollster_error, pollster_std	5	25	0.605	0.571	0.571
Samplesize, pollster_error, pollster_std	12	10	0.719	0.528	0.528
Samplesize, pollster_error, pollster_std	12	25	0.694	0.537	0.537
Samplesize, pollster_error, pollster_std, state, partisan	5	10	0.620	0.626	0.606

Independent Variables	Maximum depth	Minimum node size	Training accuracy	Test accuracy	F1-score
Samplesize, pollster_error, pollster_std, state, partisan	5	25	0.622	0.616	0.603
Samplesize, pollster_error, pollster_std, state, partisan	12	10	0.725	0.608	0.597
Samplesize, pollster_error, pollster_std, state, partisan	12	25	0.701	0.614	0.585
Samplesize, pollster_error, polldistance, state, partisan	5	10	0.627	0.608	0.581
Samplesize, pollster_error, polldistance, state, partisan	5	25	0.625	0.619	0.600
Samplesize, pollster_error, polldistance, state, partisan	12	10	0.724	0.609	0.591
Samplesize, pollster_error, polldistance, state, partisan	12	25	0.700	0.618	0.594

Table three: Testing random forest classification models to classify 4% error.

The best input variables, sample size, pollster error and standard deviation, partisan race and state, returned an f1 score of .606.

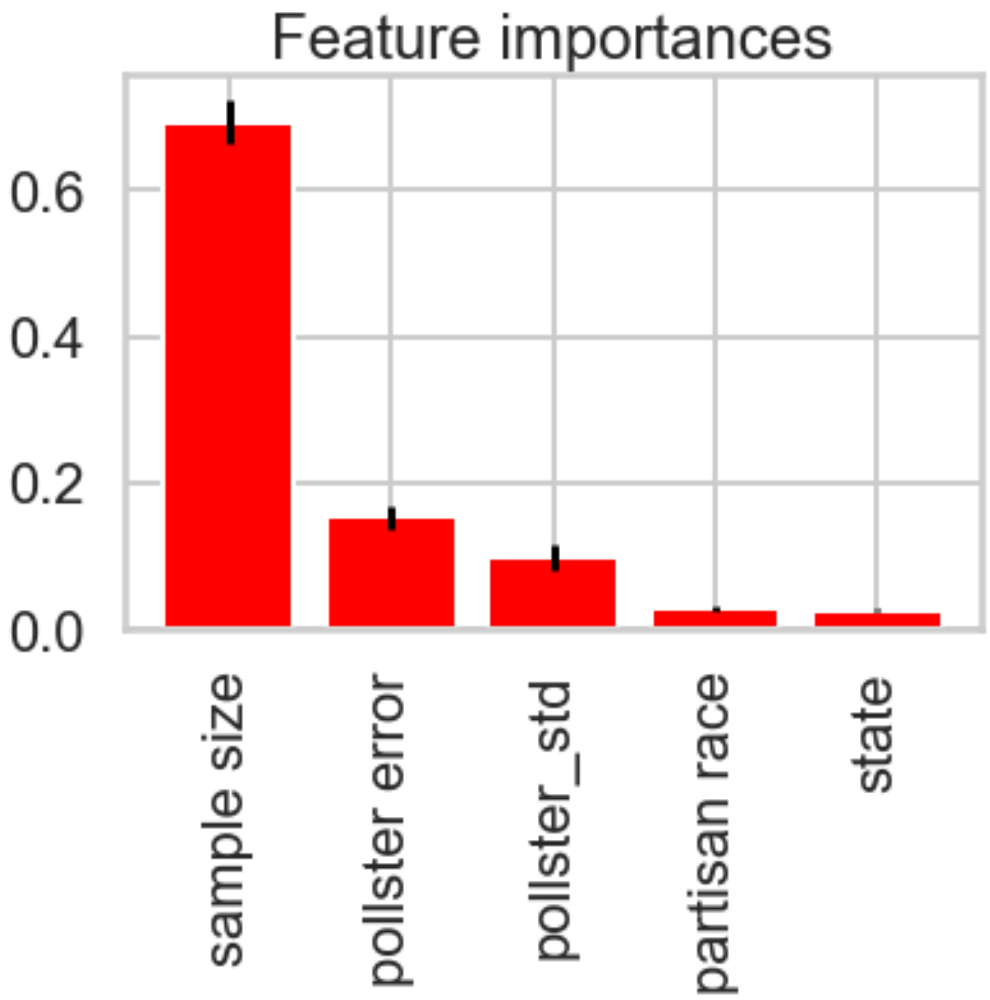

Figure Sixteen: Feature importance graph of the final random forest classification model.

We can see that sample size, pollster error, and pollster standard deviation are the primary predictors of a poll's accuracy. Both the f1 scores of models lacking partisan and state tracking and their low feature importance indicate that these features make only minor contributions to the model. Incidentally, the intermediate models for right_call classification weighted sample size at 0.6, pollster error at 0.38, and nothing else above 0.01.

4.3 Random Forest Regression

Classifying a poll as being within a margin of error is better than predicting maximum error, but we have the opportunity to go further still. With a random forest regression model, we can attempt to predict the exact error of a poll based on the same variables we've previously used.

Random forest regression models are adjusted in the same manner as random forest classification models, with maximum depth, number of trees, and minimum node size being the most relevant variables to manipulate. The difference is that the dependent variable is continuous, not binary, and the available scoring methods are different. For this reason, we pass the model the precise error of the polls for training, not the four percentage point margin used for the random forest classifier.

A f1 score, for example, is only valid for a classification model. For this reason, the random forest regression models were scored with R-squared. We can see the effect of varying maximum depth and minimum node size here:

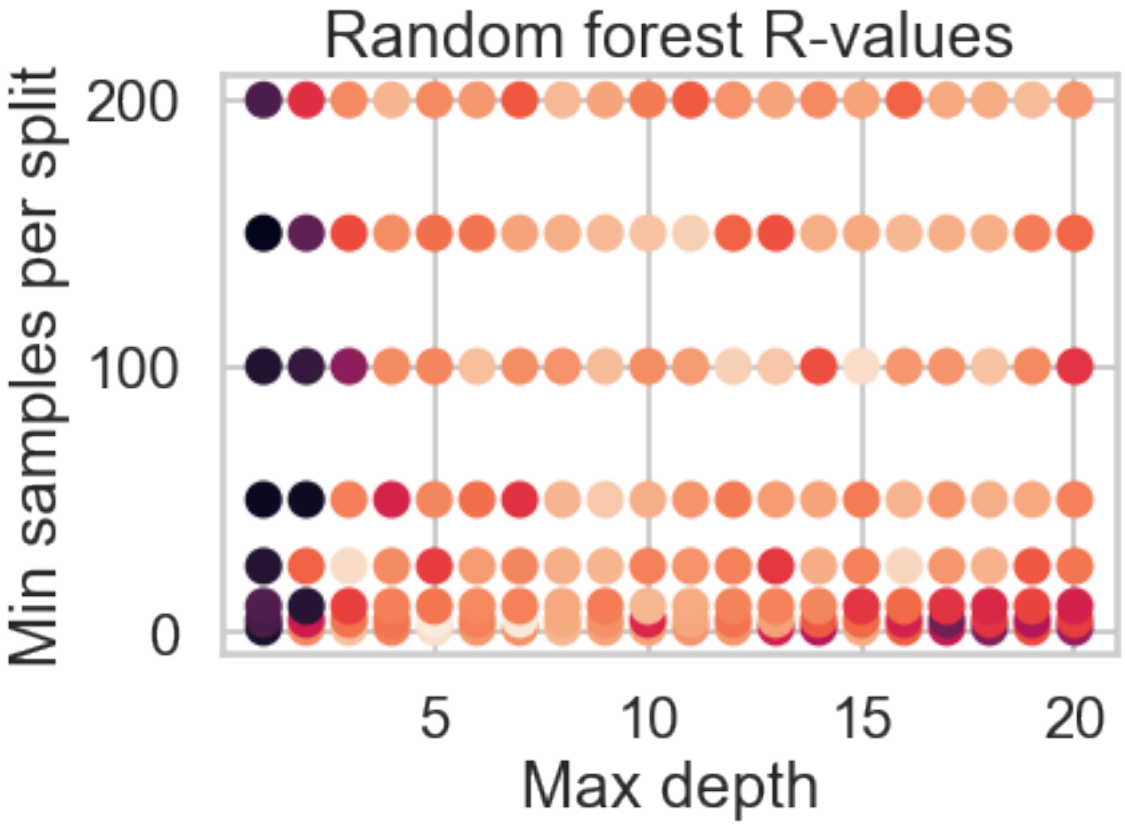

Figure Seventeen: Scatter plot of varying R based on depth and sample size.

Holding maximum depth to around five appears to again be the limit for increasing the accuracy of the model: past this point, we overfit with no benefit to the final results.

In essence, we are here attempting to predict the final outcome of a race based on a single poll and our knowledge about the pollster. Needless to say, this verges on the impossible. Our best model was only able to predict around 20% of total error – a poor showing. Ultimately,

polling data is too inexact for a model to be able to make concrete predictions about the eventual victor any more than polling organizations can.

By aggregating this data across multiple polls, a clearer picture may be possible.

5. Usage, interested parties, and visualization

Although politics affects all Americans, an in-depth analysis of the quality of polls is unlikely to appeal to the general public. These are the groups we believe will be most interested in this modeling work:

5.1. Polling Groups

Being able to predict in advance if a poll will generate an accurate result is of interest to a polling organization that needs to determine what kind of polls they should run. This model has helped to sift out concrete insights, such as ‘sample sizes above 4000 don’t have a significant effect on error and are a waste of money’, and ‘mail polls in Wyoming tend to severely overrate the Republican candidate’.

5.2. Political Candidates

Anyone running for an elected position, be it House, Senate, governor, or president, benefits from a clear understanding of their position in the race and their odds of winning. A candidate who is already in the lead might be better served playing it safe, where a lagging candidate might try to capture attention and take risks to win over voters. How much did last week’s scandal affect the polls? How much do voters care about this issue or that issue? Everything relies on accurate data, and this model helps distinguish accurate data from inaccurate data.

5.3. OLS or Random Forest?

The OLS is a simple model that gives an accurate prediction of maximum possible error. Although it conveys less information than the other models, it is the most accurate model created here. Use to establish or predict a basic margin of error for casual use. It requires only the identity of the pollster and the sample size of the poll, making it ideal for quick predictions.

The Naïve Bayes model was never intended for heavy use. It was used to better determine which variables were relevant for predicting error, but should not be used to make those predictions itself.

The random forest classification model is more complex and require a broader spread of information about the poll: the state the poll was conducted in and if the race was partisan. Because the model is attempting a more precise forecast, it is less reliable, but can still give a basic evaluation of poll trustworthiness.

The random forest regression model was unable to make sufficiently good predictions about error to be worth including, and should be discarded.

In most circumstances, it would be best to use both the random forest classification model and the OLS maximum error model to get the most information possible.

5.4 Understanding the results

It is likely to be the case that the best data to return is simply the result of the model, without significant explanation of the underlying mechanics. Consider this decision tree from a random forest model:

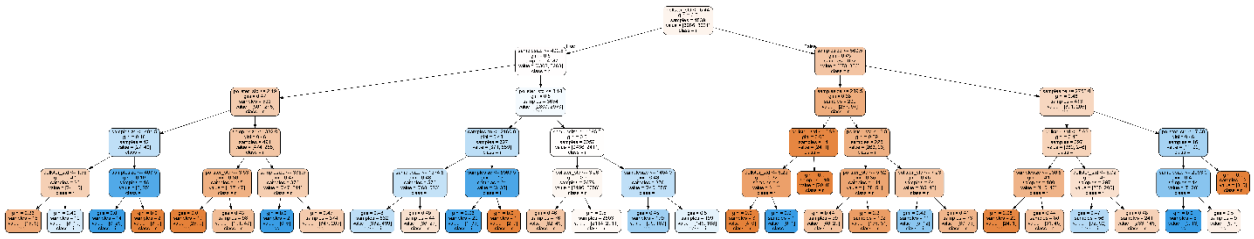

Figure Eighteen: Untrimmed decision tree, restricted to depth of five.

This is a decision tree that has been restricted to a maximum depth of five, trained on only a handful of variables, and it is still difficult to parse out how the model reaches its conclusion. More than that, it is irrelevant: Political candidates aren't likely to be interested in why they should trust a poll, only if it is trustworthy. Pollsters would be interested in which polls the model predicts to be good or bad, but that would be better conveyed by a human with knowledge of the model.

6. Limitations, Failure points, and Inconsistencies

As discussed earlier, the model considers the past performance of a pollster when determining if a new poll is trustworthy. For that reason, new pollsters will be harder to predict than established pollsters. For the same reason, if a polling company is lapsing and releasing worse and worse polls, or improving and releasing better and better ones, the model will lag behind a model that only considers statistics about individual polls, not the pollsters producing them. This is an acceptable trade-off, given that prior pollster performance is a strong predictor of future work, but these limitations should still be noted.

For the random forest model, the binary state classification and poll-distance classification mean that the model will predict artificially sharp differences between data points on the classification boundaries. The OLS model performs well within the context of its expectations, but predicting maximum error is simply less valuable than predicting precise error.

The available literature concerning the costs of polling is inconsistent and varies greatly by polling organization, state, and specific methodology – it is not enough to simply attach a price tag to online, IVR, or in-person polls. Recommending the most cost-effective poll requires input from a specific polling organization about their costs of operation, and cannot be generalized across the entire industry.

If more data about how pollsters weight their polls could be obtained, this model could be substantially improved. Pollster identity is a related variable, but not a perfect match.

7. Further development and refinement

As new polls are published, they can be added to the dataset that this model draws from. This will improve the model and help track future trends in polling. To appeal more to polling organizations, information such as response rate to different poll methodologies and cost per person polled are necessary. This would allow for a proper optimization model that calculates how to maximize respondents while minimizing cost, taking into account that not all poll methodologies are equally accurate.

Another way to improve the model's performance would be to train it on falsified and low-quality polls. The FiveThirtyEight curated dataset allowed a tighter focus, but being able to distinguish polls with heavy herding, falsified results, and extremely inaccurate projections would give the model broader appeal.

Finally, it would be possible to improve the model by further streamlining the process of taking in new data and returning predictive output, perhaps by setting up a webpage. The current interactive, a Jupyter notebook, is serviceable but clunky.

8. Final thoughts

In this project, we have examined four predictive models aimed at forecasting election results and poll quality. Maximum and exact error were found to depend heavily on the total number of people surveyed for the poll and lightly on prior performance by the pollster, both mean accuracy and consistency. The location the poll and time before election marginally improved the model, but sample size and past performance were the primary variables. Predicting election results based on this information was attempted but ultimately unsuccessful.

Sample size was the single most predictive variable examined. The simplest way to improve accuracy in polling is to reach out to more people. After sample size, the next most predictive variables are the mean error in polls previously conducted by the pollster in question and the standard deviation in said error. Pollsters that have conducted good polls in the past tend to continue conducting good polls. This is everything you need to know to make a good estimate of maximum possible error for a given poll. In contrast, time before election, state the poll was conducted in, and partisanship of the race contributed minimally to the final model.

Polling methodology, type of race, and predictive margin had no appreciable effect. We've seen that polling methodology does predict bias, but it was not worth factoring in when modeling error. Type of race and predictive margin indicate that polling methods do not suffer from reduced accuracy when a race is especially close or low-stakes.

The models we have built here all attempt to predict factors of single polls. The best way to move forward would be to aggregate these predictions to better forecast final election results. Although we have shown that these models perform quite poorly when attempting to predict results based on individual polls, attempting to forecast by election would allow us to better handle outliers and track trends. We leave this track open for further exploration at a later date.