**Springboard Data Science Intensive Course**

**Capstone Project 1**

**Modeling Polling Accuracy, from Governor to President**

**Consolidated Report**

Marcus Bamberger

**Table of Contents**

Introduction 3

Finding Data 3

Statistical Data Analysis 4

Sample Size 5

Poll Distance 6

Poll Type 7

Prior Performance 7

Modeling 8

OLS 8

Categorical Random Forest 10

Limitations, Failure Points, and Inconsistencies 11

Further Development and Refinement 12

Final Thoughts 12

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.

We have constructed a linear regression model forecasting maximum possible error and a random forest classifier predicting if a poll correctly called the election results. This is of interest to any political candidate, whether they are running for a governorship, House, Senate, or the presidency. These models will help a campaign predict if they need to aggressively win voters or focus on maintaining their support base, better understand how recent effects have shaped their popularity, and decide on a course of action to take.

1. **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](https://github.com/fivethirtyeight/data/tree/master/polls). 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.

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 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.

After cleaning and organizing, the primary dataset contains 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.

1. **Statistical Data Analysis**

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:

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.

* 1. **Sample Size**

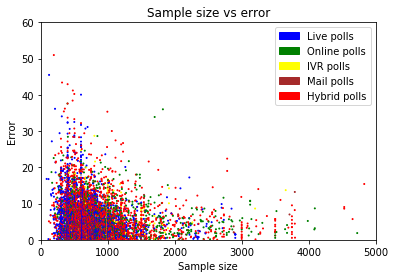


Figure One: Scatter plot of sample size versus polling error, colored by poll type.

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.

* 1. **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.

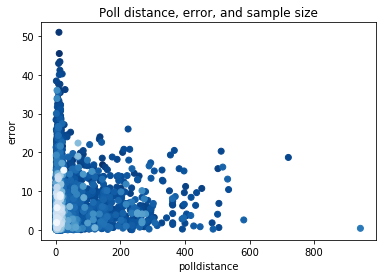


Figure Two: Scatter plot of poll distance against polling error, colored by sample size.

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.

* 1. **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?

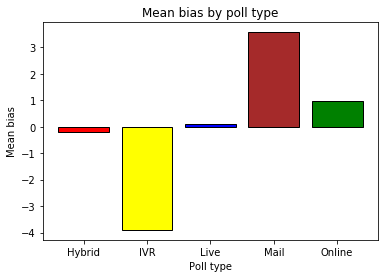
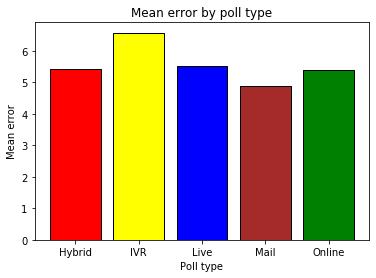


Figure Three: Mean error and bias by poll type.

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.

* 1. **Prior Performance**

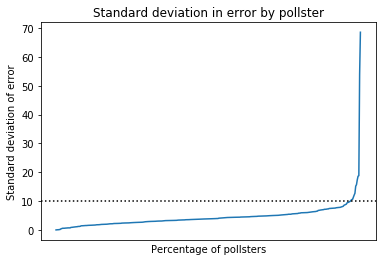
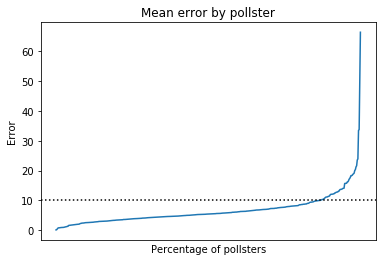


Figure Four: Mean error and standard deviation of error per pollster.

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.

1. **Modeling**
   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.

A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

Figure Five: 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.

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.

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.

* 1. **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. Gini impurity proved to be a better decision factor than entropy, and tree depth was restricted to 5-12 to minimize overfit.

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.

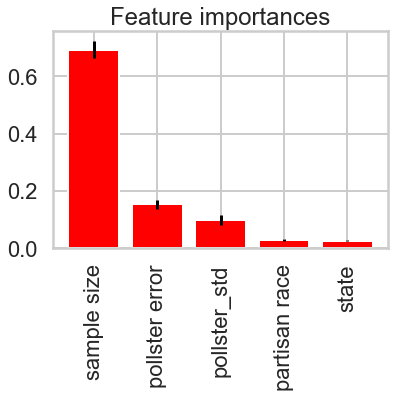


Figure Six: Feature importance graph of the final random forest classification model.

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.

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 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. 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.

1. **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.

1. **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.

1. **Final thoughts**

In this project, we have produced two 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.

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. For more information on the variables examined and the model performance, please examine our full report on the FiveThirtyEight data.

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 aggregate forecasts by election would allow us to better handle outliers and track trends. We leave this track open for further exploration at a later date.