**Abstract**

This project was intended to utilize several different public polls, and add a weight to each poll based on how well they will predict the approval rating of the President. However, given the limited amount of time, it became easier to calculate the accuracy of each poll in relation to their actual results.

**Data**

For our dataset (2016 election), there were over 26 features of the data set, which were cycle, branch, type, matchup, forecastdate, state, startDate, endDate, pollster, grade, sampleSize, population, poll\_wt, rawpoll\_clinton, rawpoll\_johnson, rawpoll\_trump, rawpoll\_mcmullin, adjpoll\_clinton, adjpoll\_johnson, adjpoll\_trump, adjpoll\_mcmullin,url, poll\_id, question\_id, createddate, timestamp, and multiversions. Before using logistic regression, two columns were added to the 2016 election csv file manually, which represented the actual result of Donald Trump and Hillary Clinton in each state of the 2016 election. This allowed us to determine whether the polls were correct or incorrect in their predictions.

In order for our logistic regression to work, we initially dropped 24 columns of the data set since they were either not numbers or were irrelevant to poll accuracy. Then, we created three columns, which were named pred\_trump, pred\_clinton, and correctResult. The final number of columns were 7, which were *rawpoll\_trump, rawpoll\_clinton, pred\_trump, pred\_clinton, actual\_trump, actual\_clinton, and correctResult*. Pred\_trump, pred\_clinton, and correctResult were binary data types (0 or 1), which allowed for the logistic regression algorithm to work successfully.

**Algorithm and Why**

The machine learning algorithm we chose to use was the Logistic Regression Algorithm. The reason why we chose the Logistic Regression algorithm was this problem was a classification problem. Using the decision boundary, we were able to make this problem into a classification problem instead of a regression one. The logistic regression worked well for this project due to the results not being linear. The Ogive part of logistic regression also allowed us to make the classification part of the project possible.

**Problems**

There were a couple of problems that were encountered with the project. First, it took a while for the logistic regression algorithm to work correctly. The pandas library presented some challenges in terms of cleaning the data set and dropping unnecessary columns. The second problem was the pandemic, where group collaboration was more difficult since everything was online.

**Final Results**

Some of the historically reliable polls did very poorly for the 2016 election according to the logistic regression algorithm. For instance, Marist College, which was also graded highly by FiveThirtyEight, barely had over 50% accuracy for Donald Trump. The two polls that did well were IPSOS poll and Morning Consult google consumer surveys. Both polls covered about half of the data from the csv file and both were over 75% accurate for predicting the poll accuracy for Donald Trump. We did not have enough time to test the results for Hillary Clinton.

**Conclusion**

In this machine learning project, we found that certain polls were more accurate than other ones, such as ABC and American Research group being 100% accurate during the 2016 Election. Also CVOTER was 92% accurate with a fairly large sample size. In future elections this project should get easier to handle because election data becomes more accessible as time passes. These algorithms could have also been used to weight each poll and predict the future approval rating of The President of The United States.

**Links**

<https://github.com/ClintWyatt/Poll-Predictor>