Polling Data Empirical Error Modeling

Data Science Institute
COLUMBIA UNIVERSITY

Yingxin Zhang, Jiaying Zhang, Jingyu Ren, Qinwei Zhao
David Rothschild (Industry Mentor)
Andrew Gelman (Faculty Mentor)

Data Science Capstone Project with Microsoft Research

Project Description

Public opinion polls have long been used to predict election results. Traditional probability-based sampling methods, such as random-digit dialing, are slow and expensive. To compare with, non-probability based sampling methods are proved to be faster, cheaper, and perhaps more accurate way to collect public opinion polls.

Building on to the belief that total survey error in election polls has always been understated, the goal of this project is to figure out the smallest sample size that can minimize total survey error, therefore derives a cheaper, more accurate, and more efficient polling method.

Data and Methodology

The data we used was a population level dataset taken from voter file and census, provided by Microsoft Research. It had total of 9000+ rows and contained demographic information such as party preference, gender, age, race, education, etc..

Our methodology is to fit different classification models with different sample sizes drawing from simple random sampling. The best model would result in a smallest sample size that makes sample coefficients converge to population coefficients.

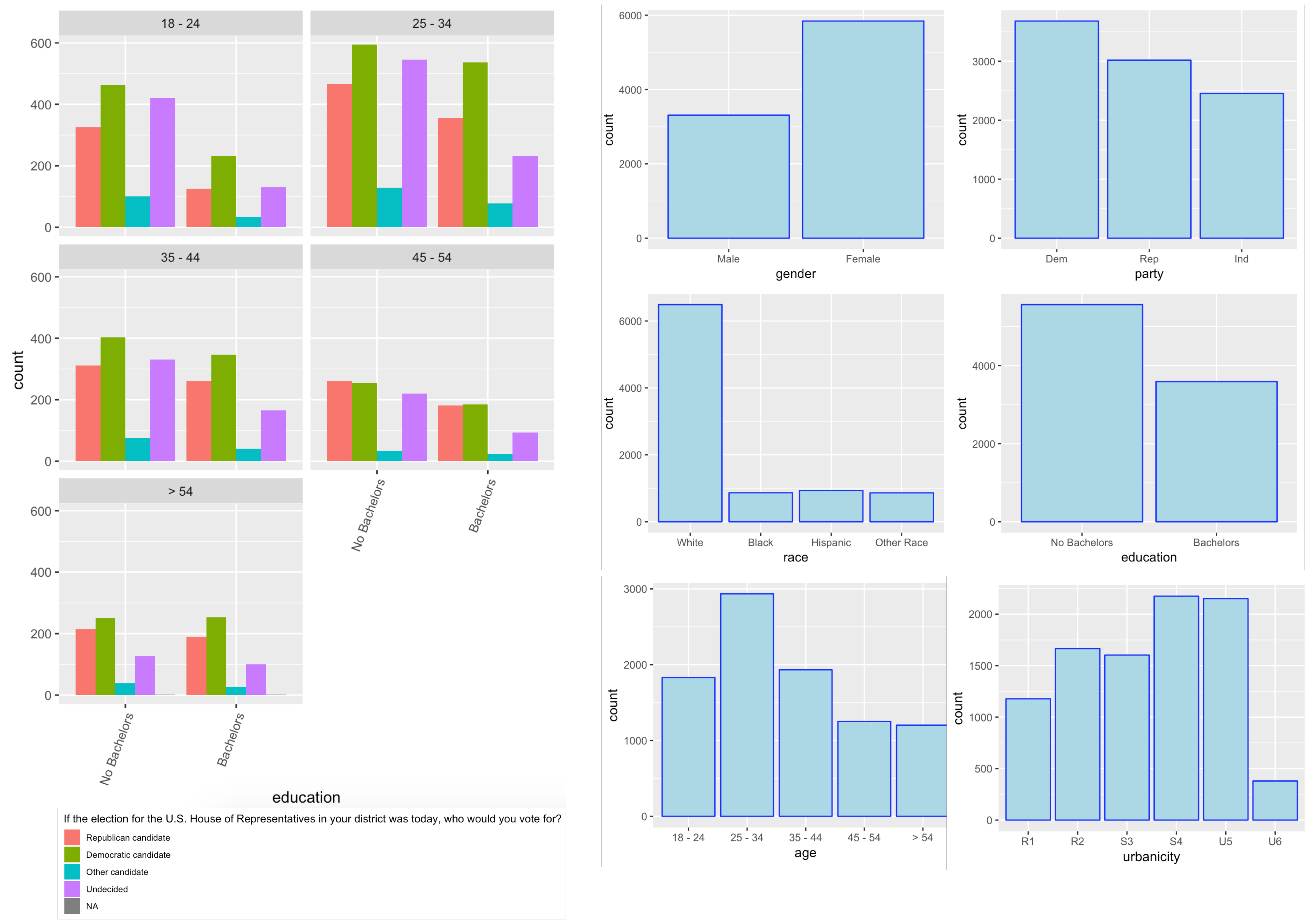

Figure 1. Relationships between parties' voting counts, voter's age and education background (left); Independent relationship between total voting counts with gender, party, race, education, age and urbanicity (right).

Results

For each sample size, we ran the following models 100 times: Random Forest, SDG Classifier, KNN, SVM Classifier, and XGBoost. The graphs below show the average output coefficients for SVM Classifier, which has the high prediction accuracy on the hold-out set, and its corresponding 95% confidence interval. The horizontal line near 0 represents the coefficient value for the entire dataset. As we can see from the graphs , sample coefficients starting to converge to population coefficient roughly at sample sizes of 500-600.

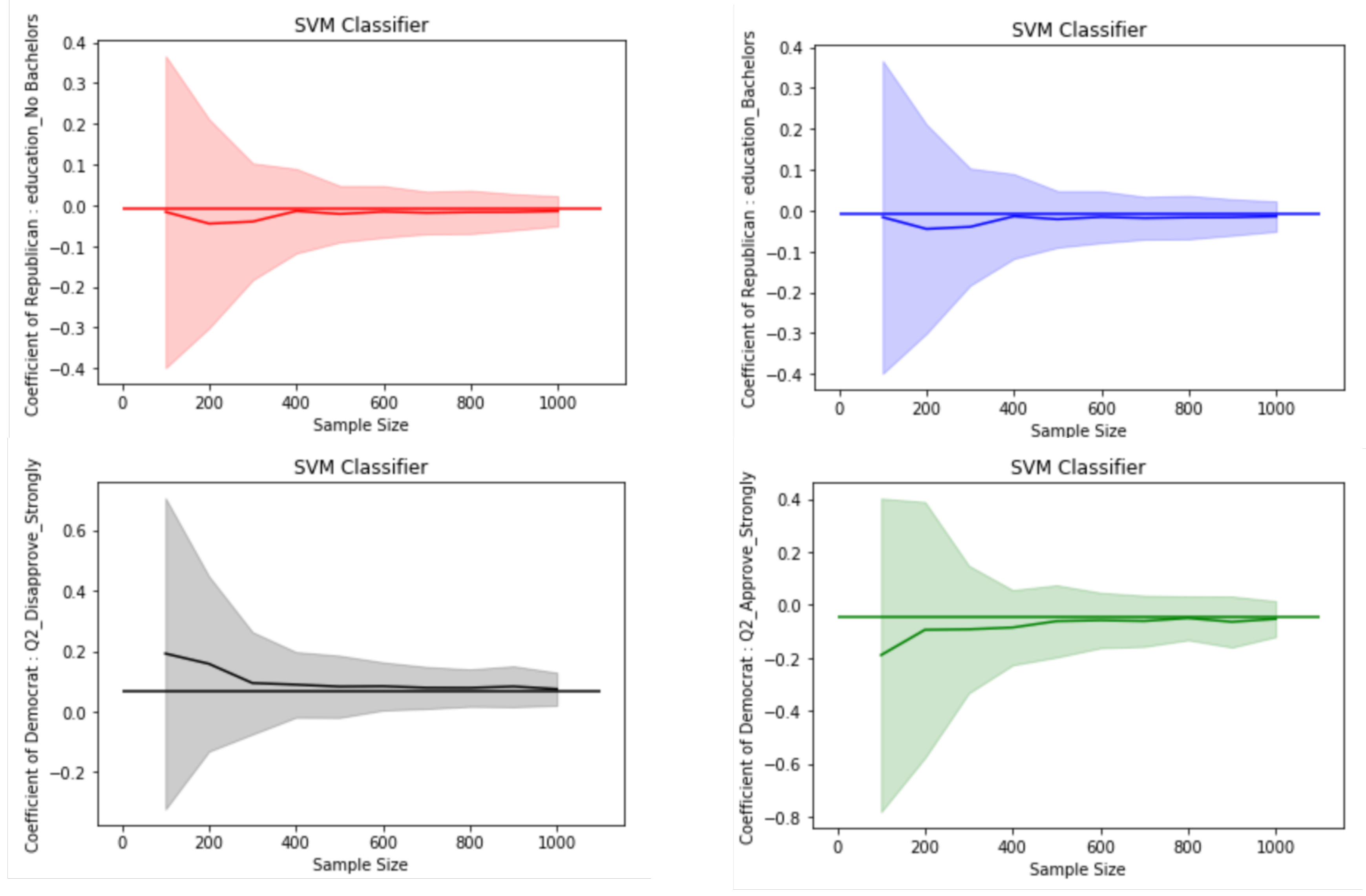

Figure 2. Mean coefficient plots for SVM Classifier with 95% confidence interval

Conclusions

This experiment shows that with much smaller size, we can still achieve the same prediction accuracy as the population level prediction. In this way, we can reduce 92% of the data collected and can significantly lower the cost and time. To obtain a higher prediction result and lower the cost, future work can be done on designing polling questions.

Acknowledgments

We would like to show our gratitude to Columbia Data Science Institute for giving us the opportunity for this capstone project. The project was supported by Dr. David Rothschild from Microsoft Research and Professor Andrew Gelman from Columbia University. We thank Tobias Konitzer, C.S.O and Co-founder of PredictWise who provided insight and expertise that greatly assisted the project.

References

Houshmand Shirani-Mehr, David Rothschild, Sharad Goel & Andrew Gelman (2018) Disentangling Bias and Variance in Election Polls, Journal of the American Statistical Association, 113:552, 607-614, DOI: 10.1080/01621459.2018.1448823