bluelabs

COLUMBIA UNIVERSITY | QMSS
QUANTITATIVE METHODS IN THE SOCIAL SCIENCES
MASTERS PROGRAM IN THE INSTITUTE FOR SOCIAL AND ECONOMIC RESEARCH AND POLICY

Examination of Sample Size Variation and Model Accuracy

Bluelabs Analytics
QMSS Fall 2022 Practicum Project

bluelabs

COLUMBIA UNIVERSITY | QMSS
QUANTITATIVE METHODS IN THE SOCIAL SCIENCES
MASTERS PROGRAM IN THE INSTITUTE FOR SOCIAL AND ECONOMIC RESEARCH AND POLICY

Contents

bluelabs

Introduction & Objectives

Build an end-product that can accurately weigh the trade off between sample size and model accuracy

01 Build and refine support models for three types of elections:
- 2020 Presidential Election (one model each for non-reg and reg states)
- 2020 Kentucky Senate Election
- 2021 Virginia Gubernatorial Election

02 Validate models to ensure predictive accuracy on topline sample as well as key groups

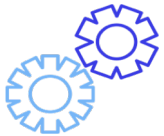
03 Conduct sample analysis on above models to determine variation of accuracy rate dependent on n size

bluelabs

Methodology

Data Processing

- Initial cleaning of data to remove unnecessary data, impute values, etc.
- Subsetting data based upon dependent variables
- Recoding of categorical variables and removal of others to address issues of multicollinearity, etc.



Modelling Approach

- Created 4 logistic support models, one for each set of election data
- Began feature selection process using supervised L1 regression to identify significant features
- After manual review and adjustment, imputed these features into Logistic Support Model to form predictions on training, testing, and full data separately

---

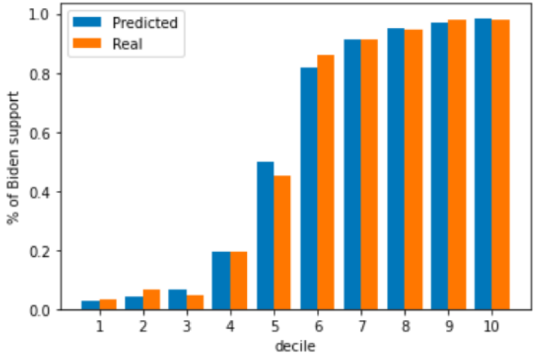
Monte Carlo Simulation

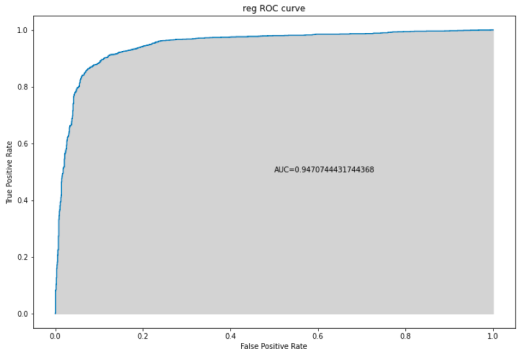
- Define a relevant series of sample sizes for each of the four models
- For each permutation of sample size:
  - Generate pulls from the sample population using a stepwise function until maximum n size is reached
  - For each subset of the sample, control sample population using random seed from 1 to 100
  - Run relevant logistic support model on each sample
- Calculate summary statistics and plot ROC score for each sample size permutation

bluelabs

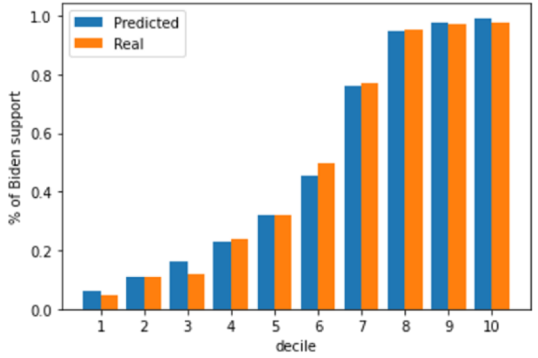
Model Validation

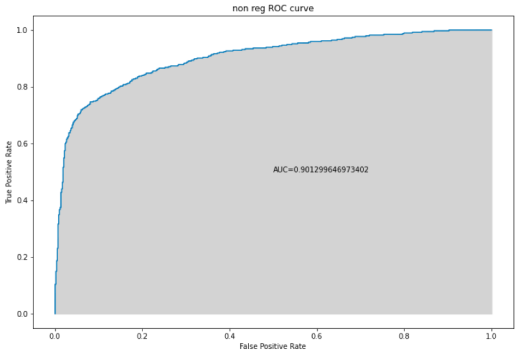
Presidential (Reg)





Presidential (Non-Reg)

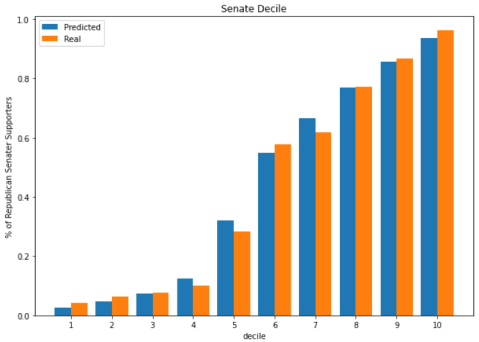


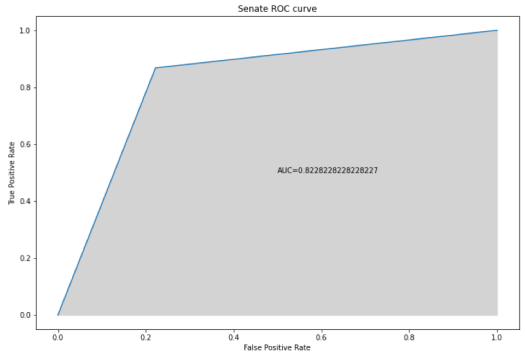


bluelabs

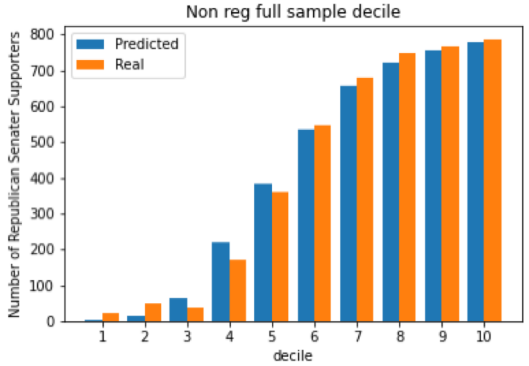
Model Validation (cont.)

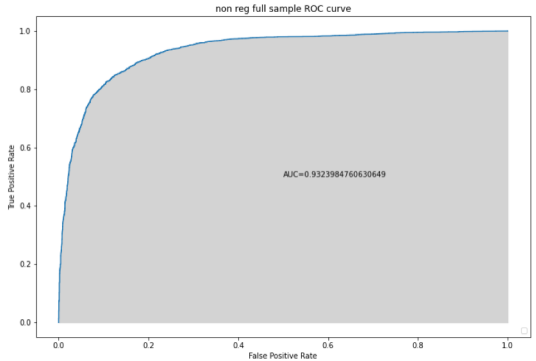
Senate





Governor



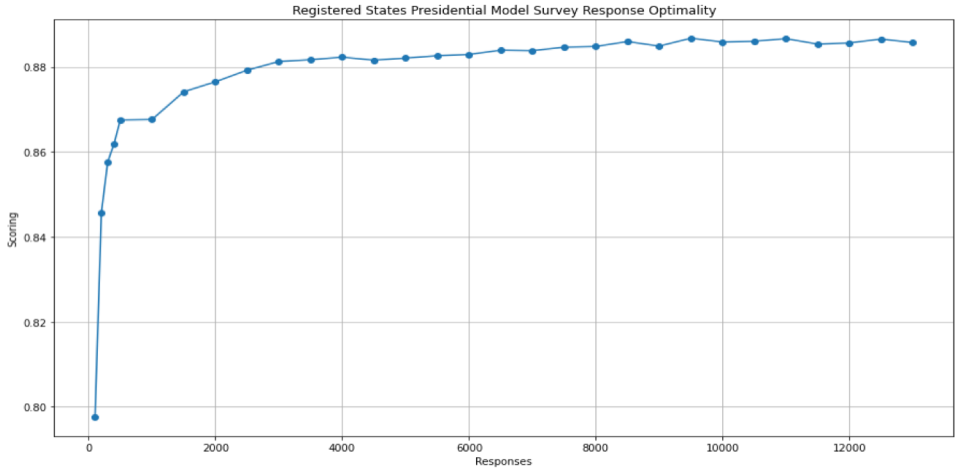


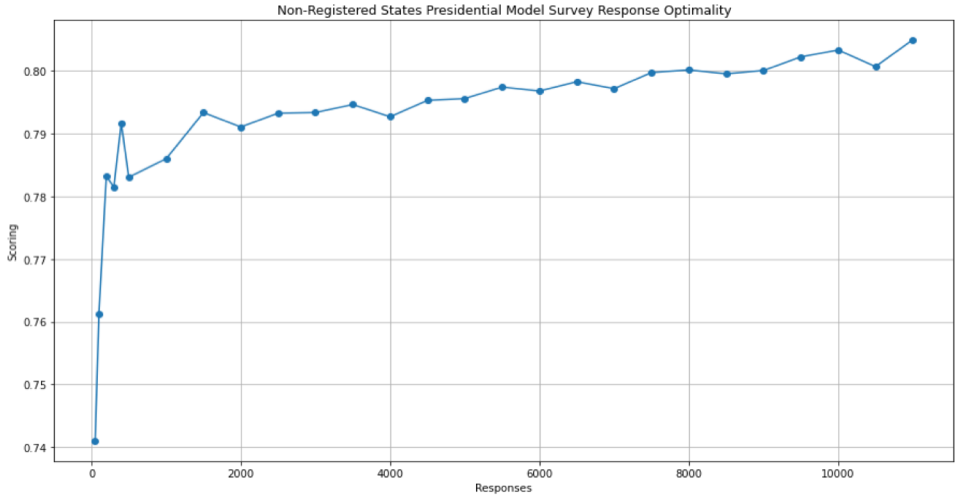
bluelabs

Findings:

Presidential Team

- Registered States
  - Accuracy rate tends to reach a horizontal asymptote at ~90%.
  - The marginal utility of increasing n size begins to diminish significantly after a sample size of 2000.
  - An accuracy rate of 85% is possible with a sample as small as 1000.
- Non-Registered States
  - Accuracy rate for non-registered states tends to be lower than registered states, with an upper limit of ~83%.
  - A similar pattern of decreasing marginal utility at the 2000 mark is also apparent.
  - An accuracy rate of 79% is possible with a sample of 2000 responses.



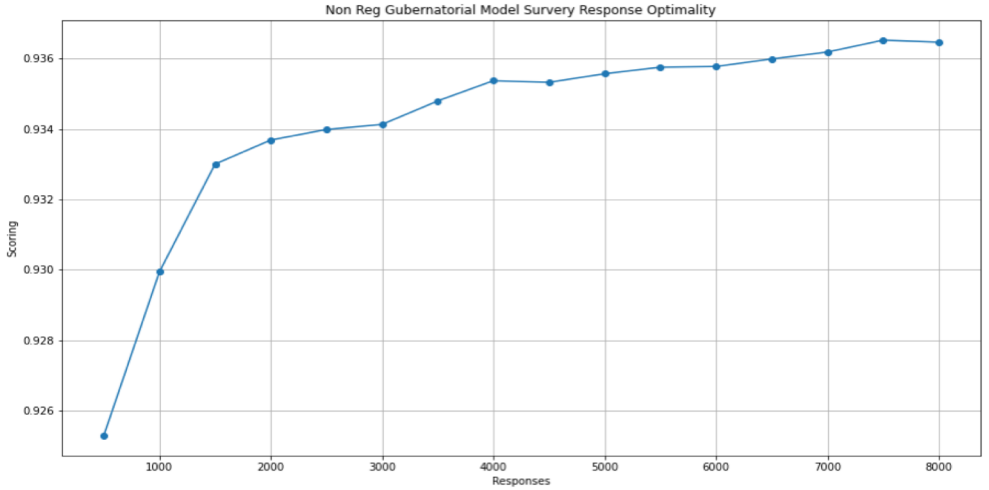


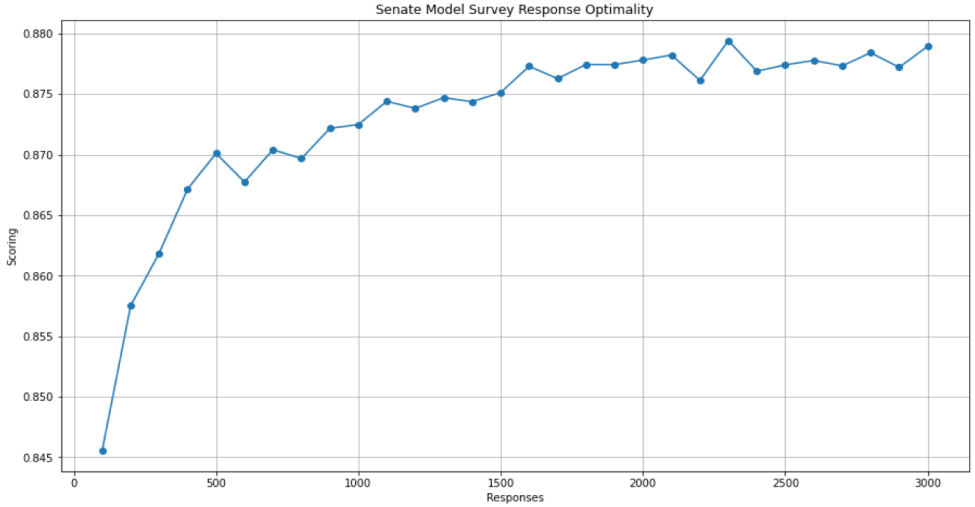
bluelabs

Findings:

Senate and Gubernatorial Teams

- Governor Model
  - Accuracy rate of near 88% reached at an n size of ~2300.
  - Accuracy of 93% possible with n size of 1000.
- Senate Model
  - Diminishing returns in accuracy after an n size of 4000.
  - Accuracy of 87% possible with an n size of as small as 500.





Conclusions

Results suggest that we can still get decent accuracy rates with smaller (sometimes significantly so) sample sizes. However…

- …is the same pattern repeated for accuracy at a more granular level? For example, good topline accuracy rates but perhaps less so for specific demographic groups or regions with models trained on smaller sample sizes

- …related to the above point, small sample size can exacerbate issues with regards to categories (eg. states) with overly small samples. Could MRP help ameliorate these problems?

Q & A