Polling Data Empirical Error Modeling

Final Report

Capstone Project - Microsoft Research

Abstract

Traditional probability-based sampling methods, such as random-digit-dialing (RDD), were used to gauge public opinions, but they are often slow and expensive. Non-probability based sampling methods have proved to be a faster and less expensive way to collect survey data. In this project, we explore samples of different sizes on different machine learning models, with the goal to find out how relatively a small sample size can produce similar prediction results when we use population data. We utilize polling data collected from mobile survey platform and build classification models (random forest classifier, stochastic gradient descent (SGD) classifier, k-nearest neighbors classifier (KNN), support vector machine classifier (SVM), and XGBoost) for individual-level predictive modeling. For each model, we fit 10 different sample sizes of data, each with 100 times drawn from simple random sampling. We choose the best model SVM based on comparing sample and population prediction accuracy and AUC score. We then further examine whether sample model can reproduce population model performance by looking for sample coefficients convergent to population coefficients through visualization. In the end, we find the coefficients of SVM classifier converge to population-level values when sample size is 500. Eventually, we reduce more than 90% of the original data volume collected in this example, successfully lower the cost and time spent on collecting survey responses.

Introduction

Public opinion polls have long been studied by researchers and scientists to predict election results using statistical modeling. Traditional probability-based sampling methods that were used to gauge public opinions, such as random-digit-dialing (RDD), are often slow and expensive. On the other hand, non-probability based sampling methods has proved to be a faster and less expensive way to collect survey data. It is also a well-known issue that election polls suffer from a variety of sampling and non-sampling errors, which we collectively refer to as total survey error. These underestimates will sometimes lead to serious polling failures, such as the 2016 U.S. presidential election. Hence, it is important to quantify the structure of polling errors using statistical models.

Our methodology is to fit different machine learning classification models with different sample sizes drawing from simple random sampling. Models and methods include Random Forest, Stochastic Gradient Descent (SGD) Classifier, K-Nearest Neighbors (KNN), Support Vector Machine (SVM) Classifier, and XGBoost. The best model would result in a smallest sample size that makes sample coefficients converge to population coefficients. The goal of this project is to figure out that smallest sample size that can minimize total survey error, therefore derives a cheaper, more accurate, and more efficient polling method. We will be working on individual level population data rather than state level or demographic group level data.

Data Discussion

❏ Data Description

The data we use is a population level dataset taken from voter file and census, provided by Microsoft Research. It has a total of 9000+ rows and 27 columns of data. It contains demographic information such as gender, age, race, education, and marital status, etc, as well as 16 questions related to party affiliation, policy opinions, and voting preference. Out of the 27 columns, we choose 9 features, consisting of 4 demographic information and 5 survey questions, as our training data. We have two prediction targets: 1. "If the election for the U.S. House of Representatives in your district was today, who would you vote for?"; 2. "Do you think the Republican candidate or the Democratic candidate is more beholden to special interest money from oil and gas companies?". In terms of answers, there are 4 to the first question: "Democrat candidate", "Other candidate", "Republican candidate", and Undecided"; there are 3 to the second one: "Both equally/no opinion", "Democrat candidate", "Republican candidate".

❑ Exploratory Data Analysis and Visualization

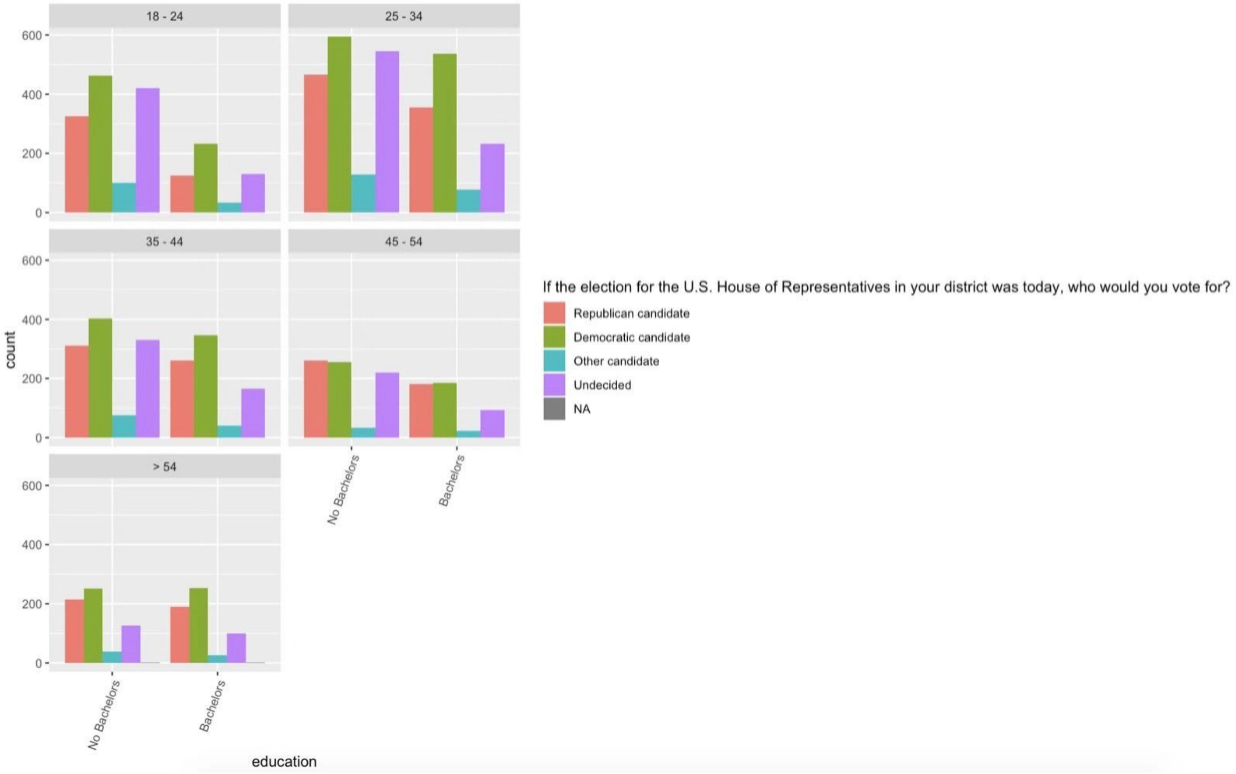


Figure 1. Distribution of answers to the question “If the election for the U.S. House of Representatives in your district was today, who would you vote for?” grouped by age and education.

In figure 1, we can see voting patterns are different in different age groups even if they have same the educational background. In all age groups except for ‘45-54’, more voters are willing to vote for democratic candidate. Republican candidate get slightly more or almost equally support in the groups of voters in age of 45-54. Another thing to point out is, voters without bachelor degree are more likely to vote for other candidate, in comparison to those with bachelor degree.

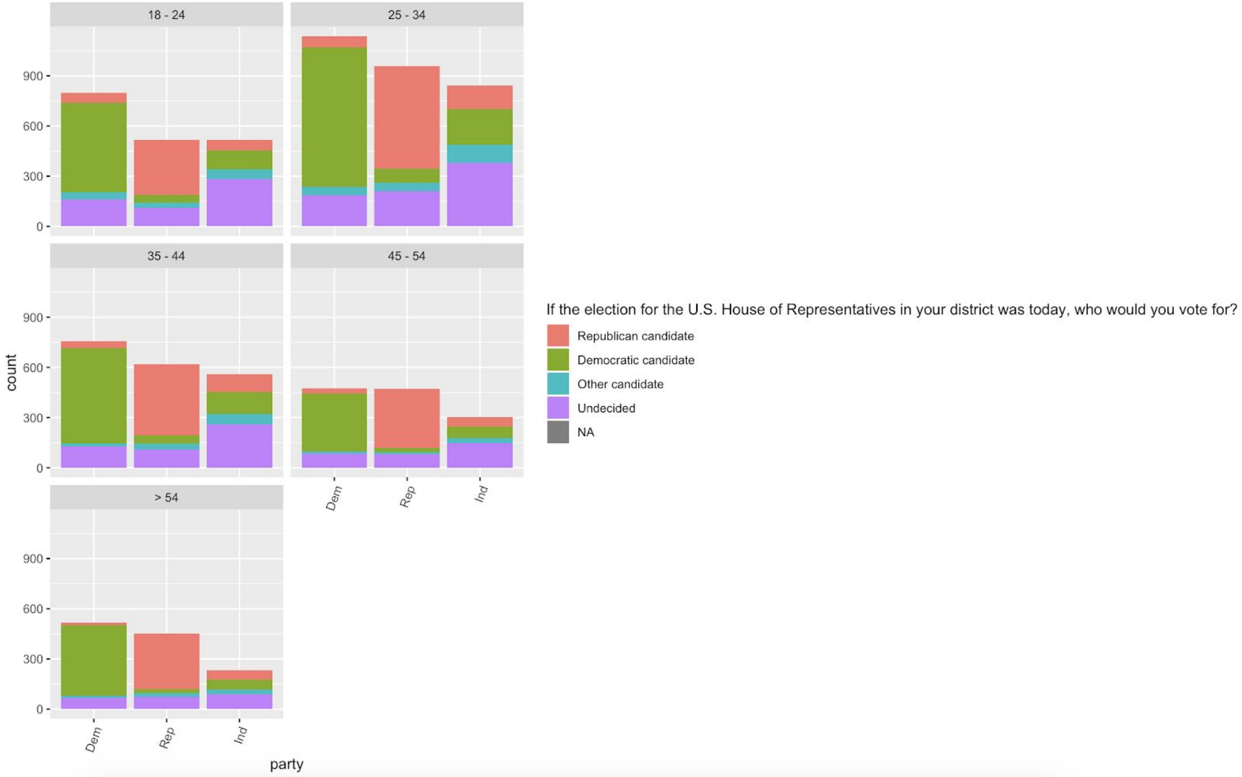


Figure 2. Distribution of answers to the question "If the election for the U.S. House of Representatives in your district was today, who would you vote for?" grouped by party and age.

We can see from figure 2 that the majority of voters will vote for their parties, i.e., people in republican party would like to vote for republican party and people in democratic party would like to vote for democratic party. People in independent party have almost equal willingness to vote for democratic or republican. Our model would run into troubles predicting overall election results when respondents are all from one of these two major parties. If respondents are all from democratic party, model would predict winning of democratic party. But this would not reflect the real election results in the future. So make sure to check the weight of democratic republican respondents before modeling.

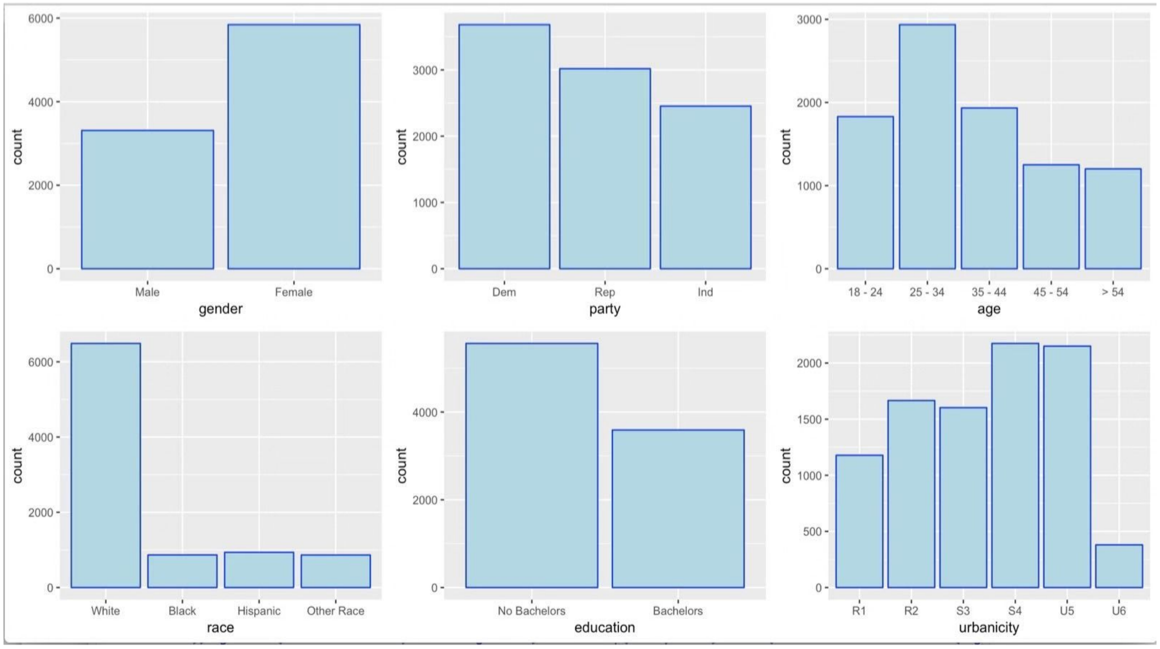


Figure 3. Distribution of different demographics

Figure 3 shows respondents' demographic distribution. From these plots we can have an overall understanding of who participated in this online polling survey and the profile of these respondents. These plots reflect following points: Graph in the upper left corner shows that proportion of male and female is about 1:2. More respondents in demographic party participated in this survey compared to those in republican and independent party. Respondents in age of '25-34' are more likely and actively to participate in online surveys and we got most responses from this group of people. About 66% of polling data was acquired from white people. More respondents did not have bachelor degree in the past. At the meantime, responses from different urbanicities were not equally distributed.

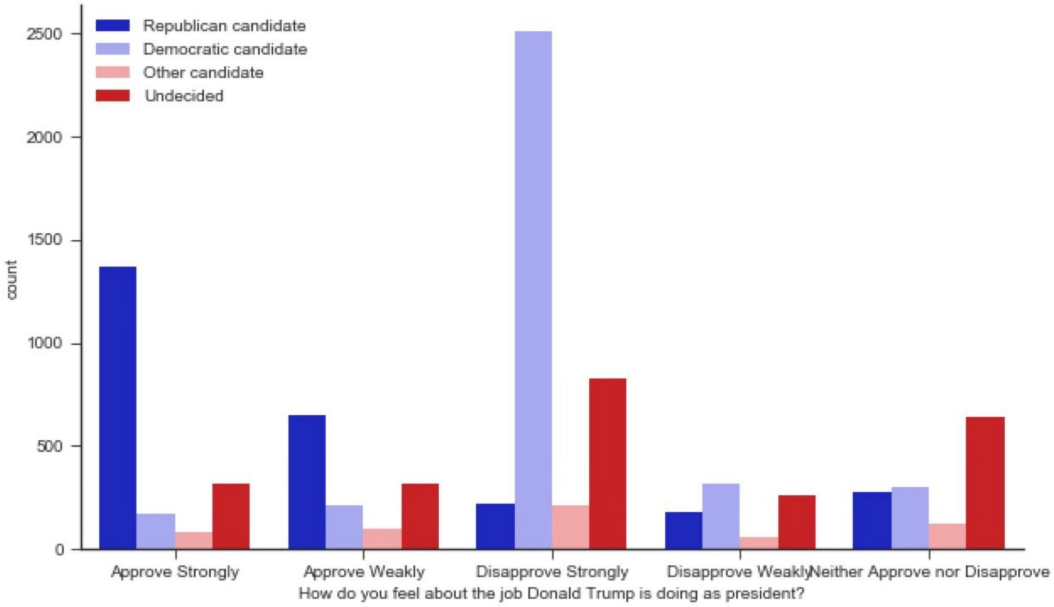


Figure 4. Distribution of answers to the question “If the election for the U.S. House of Representatives in your district was today, who would you vote for?” grouped by “How do you feel about the job Donald Trump is doing as president?”

We can see from figure 4 that voters will vote depending on their preference on Donald Trump. Specifically, voters disapprove strongly about Donald Trump will vote for democratic candidate and voters approve strongly about Donald Trump will vote for republican candidate. Only the minority voters who neither approve or disapprove decide not to vote. If the polling data is imbalanced (respondents all disapproved strongly for Donald Trump or approved strongly for Donald Trump), our prediction will not be stable when we use another polling data or future polling. So make sure to check the weight of respondents preference about Donald Trump before modeling.

These skewness and inconsistency of census data might introduce difficulties and inaccuracy when building models. For individual-level modeling, it will not cause many problems because machine learning models are robust. However, for state-level regression, it is better to re-weight our polling data and change it to census data with the same demographic distribution.

❑ Data Preprocessing

To prepare the data for model training, we first extract the 9 feature columns as a new data frame. The 9 features consist of the following: ‘education’, ‘gender’, ‘age’, ‘race’, ‘What is your political party affiliation?’, ‘How do you feel about the job Donald Trump is doing as president?’, ‘Which of these seven topics do you care about most?’, ‘How do you feel about the job the Republican candidate is doing as a member of congress?’, ‘Do you think the Republican candidate or the Democratic candidate would do a better job addressing the issue of public health?’. We also extract the two prediction target columns mentioned above. To ensure the training data are clean and formatted, we drop rows with empty entries and result in a total of 9103 rows. Since all data are categorical, we apply One-Hot-Encoding to our feature matrix. It produces a new data frame of dimension (9103, 40). We also apply label encoder to transform our target variable to numeric data. After this transformation, we use standard scaling method to scale the feature data.

Modeling Approaches and Critical Results Discussion

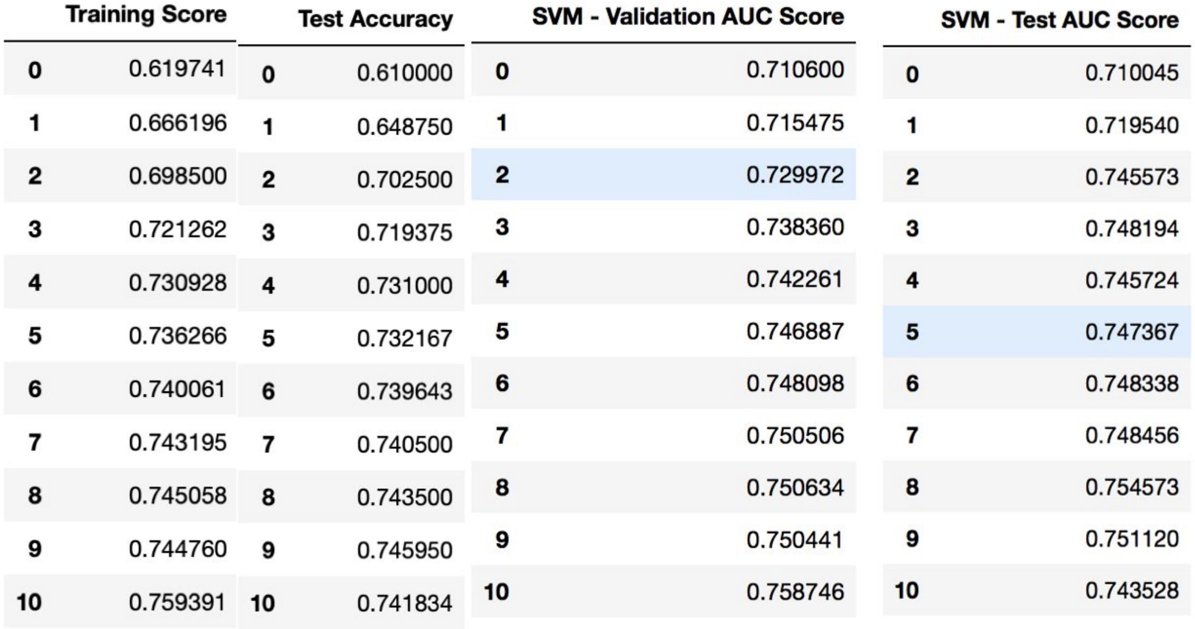
❑ Model implementation and Comparison

Since we want to check the performance of model in terms of survey error, we need to choose different sample sizes and run a model a certain amount of times to see if we can obtain sample model performance that converges to population model performance. Using the dataset and preprocessing method we introduced in the above section, we choose ten different sample sizes ranging from 100, 200, 300, …, to 1000, which are typical order of magnitudes for polling responses.

For each of our chosen sample sizes, we try running several classification models and each time with simple random sampling from the population dataset. The models we use are the following: Random Forest classifier, Logistic Regression, Stochastic Gradient Descent (SGD) Classifier, K-nearest Neighbor (KNN) Classifier with 3 neighbors and 5 neighbors, Support Vector Machine (SVM) Classifier, and

XGBoost. We train each model using a 10-fold cross validation and evaluate it on a 20% hold-out test set. We mostly utilize scikit-learn and matplotlib libraries for model training and evaluation and visualization.

Among these models, we find that SVM Classifier is the one with the highest prediction accuracy and AUC score on the hold-out set, whereas for other models, the highest score can only reach about high 60%. Figure 5 below shows training and test accuracy, validation AUC, and test AUC scores. The 0-9 rows represent sample model evaluation ranging from sample sizes of 100 to 1000. The last row represents population level model evaluation. Results show that models trained on sample data can have comparable performance to the model trained on population level data, starting from sample size of 600-700 in this case.



	Training Score	Test Accuracy	SVM - Validation AUC Score	SVM - Test AUC Score
0	0.619741	0.610000	0.710600	0.710045
1	0.666196	0.648750	0.715475	0.719540
2	0.698500	0.702500	0.729972	0.745573
3	0.721262	0.719375	0.738360	0.748194
4	0.730928	0.731000	0.742261	0.745724
5	0.736266	0.732167	0.746887	0.747367
6	0.740061	0.739643	0.748098	0.748338
7	0.743195	0.740500	0.750506	0.748456
8	0.745058	0.743500	0.750634	0.754573
9	0.744760	0.745950	0.750441	0.751120
10	0.759391	0.741834	0.758746	0.743528

Figure 5. SVM Classifier training & test accuracy, validation AUC score & test AUC score for 1st prediction target

Further step to determine whether model trained on sample data is representative enough is to compare sample model coefficients to population model coefficients, aiming to see if sample coefficients converges to population coefficients. Since we have 40 independent variables and 4 classification outcomes, we have a total of 160 coefficients every time we run the model. Out of these 160 coefficients, we randomly pick 4 to examine. These 4 coefficients are: respondents with no bachelor degree who vote for republican, respondents with a bachelor degree who vote for republican, respondents who answer "disapprove strongly" to feature question No.2[1] and vote for democrat, and respondents who answer "approve strongly" to feature question No.2 and vote for democrat.

The graphs below show the average of the 100 coefficients output and their corresponding 95% confidence interval versus different sample sizes. The horizontal line near 0 represents the population level coefficient value. As we can see from the graphs, sample coefficients start to converge to population coefficient roughly at sample sizes of 500 - 600 and stay relatively stable both on coefficient values and their confidence interval.

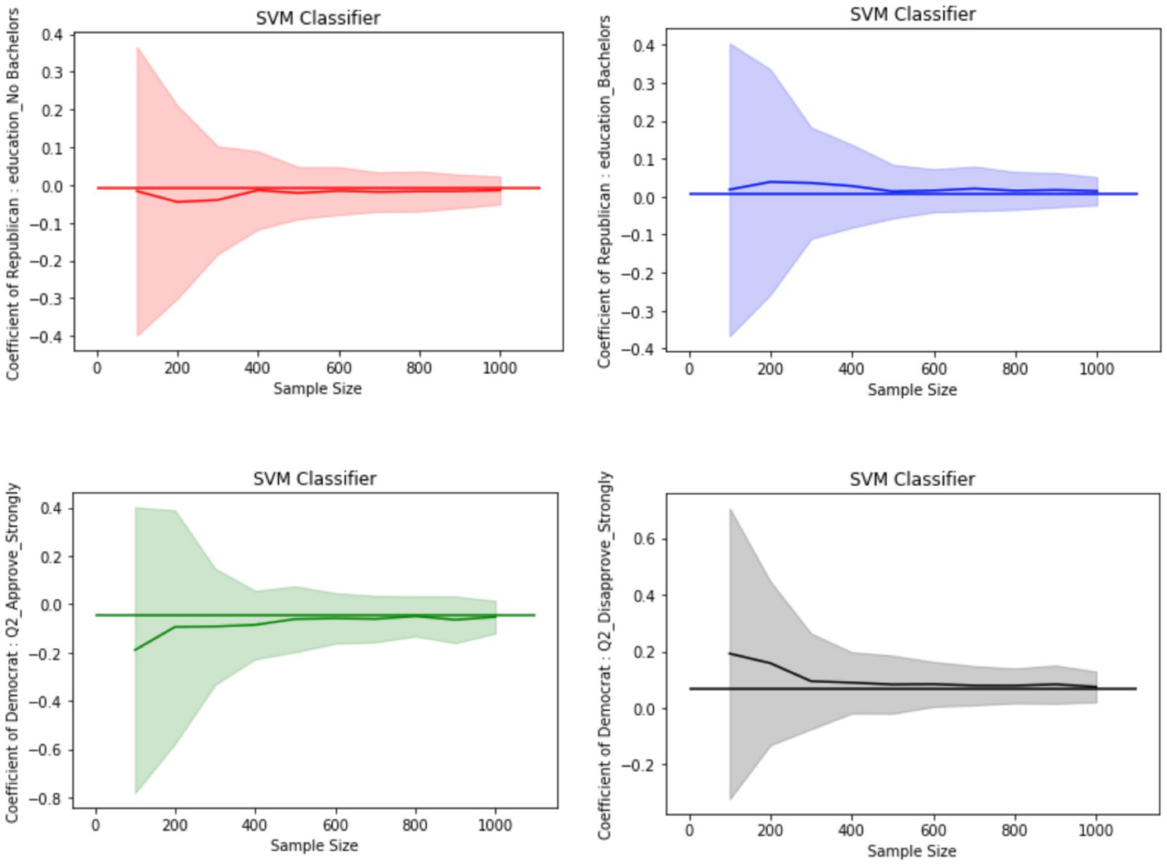
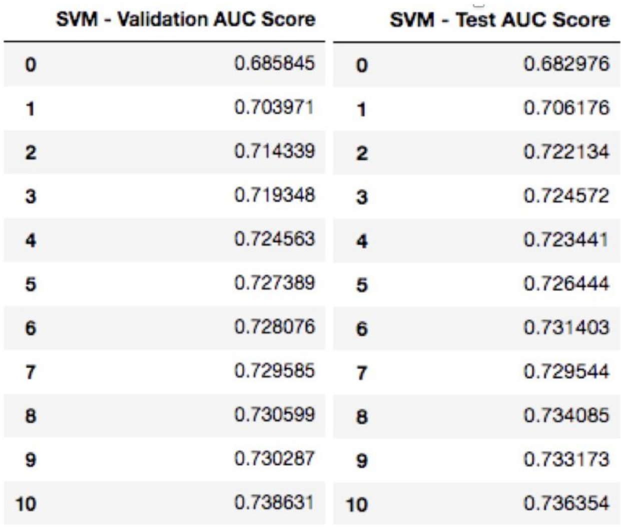


Figure 6. SVM Classifier coefficient plots for question 1

[1] Feature question No.2: How do you feel about the job Donald Trump is doing as president?

In order to examine whether our model achieves the relatively accurate and stable results is an one-off event or not, we also try another question as our prediction target using the same feature data for training. Figure 7 shows validation AUC score and test AUC score for the second prediction question. Although model performance is slightly worse than the one from trial 1, sample model performance is still comparable to population model performance, not showing significant deviations. (e.g. the highest sample AUC score is only about 0.4, while population AUC score is 0.7)



	SVM - Validation AUC Score
0	0.685845
1	0.703971
2	0.714339
3	0.719348
4	0.724563
5	0.727389
6	0.728076
7	0.729585
8	0.730599
9	0.730287
10	0.738631

	SVM - Test AUC Score
0	0.682976
1	0.706176
2	0.722134
3	0.724572
4	0.723441
5	0.726444
6	0.731403
7	0.729544
8	0.734085
9	0.733173
10	0.736354

Figure 7. SVM Classifier validation AUC score & Test AUC score for the 2nd prediction target

We also randomly pick 4 coefficients to visualize. The 4 coefficients are: respondents with no bachelor degree who choose republican candidate, respondents with a bachelor degree who choose republican candidate, respondents who answer "disapprove strongly" to feature question No.2 and choose the no opinion option, and respondents who answer "approve strongly" to feature question No.2 and choose the no opinion option.

The visualization for coefficients are shown in Figure 8 below. This time the first two coefficients achieve better convergence tendencies, starting at sample size of 300. Their 95% confidence interval also shrink significantly at sample size of 400 and keeps a much slower shrinking pace until it stays stable.

The next two coefficient plots show that sample coefficients converge to population coefficients starting from sample size of 500.

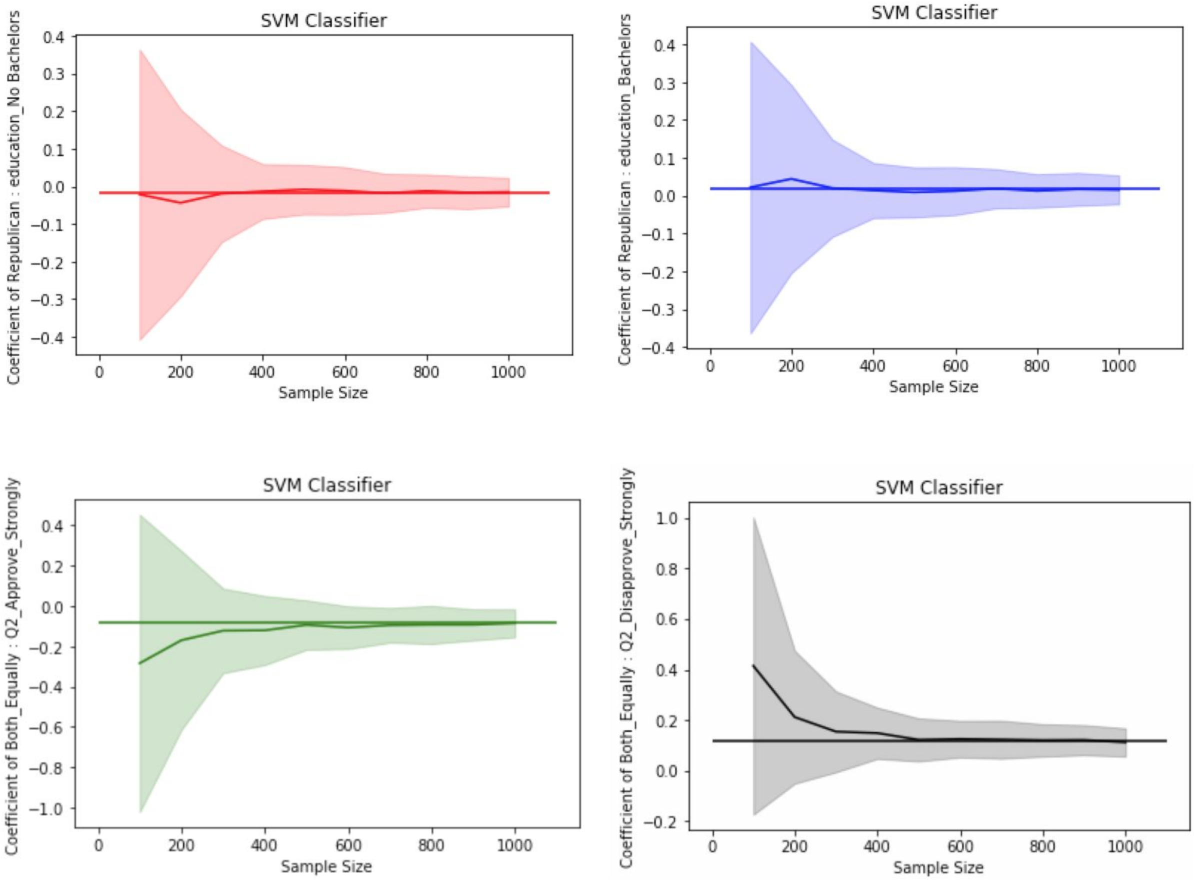


Figure 8. SVM Classifier coefficient plots for question 2

Both trails achieve almost identical results. It shows that SVM model indeed is suitable for predicting polling data at the individual level across the entire population. Although it is a little computational costly, the model performance is fairly stable and is producing satisfactory results.

Limitations, Conclusions and Future Work

❑ Limitations

The model only predicts voting pattern of potential voters, but it cannot predict who will vote and changes of voting pattern. Political campaigns have influence on people's willingness, but our model seems not focus on these changes. This report is a good starting point to try different machine learning models and gradient descent algorithm for optimization. Also we can not reach out certain groups of people who do not have access to mobile devices and the polling survey APP we used to collect responses. We can minimize modeling error using machine learning models, but we cannot capture errors that due to sampling techniques.

❑ Conclusions

This experiment shows that a small amount of sample size is enough to closely represent population in terms of election prediction. After implementing selected machine learning models on our chosen dataset, we find that SVM classifier is the one with the highest prediction accuracy among all models. Approximately 500 samples will be enough to achieve our goal. Eventually, we reduce more than 90% of the data volume collected in this example, successfully lower the cost and time spent on collecting survey responses.

❑ Future Work

First, we can use our model on another polling data, a much larger data, for example, to see whether our model still has its stability. Second, we can improve our model by using multilevel regression, building on to the belief that multilevel modeling is an improvement over classical regression. We can use data with hierarchical structure. For example, we can stratify individuals based on states or education levels.

Related Work & Member Contributions

❑ Related Work

1. 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
2. Tobias Konitzera, Sam Corbett-Daviesa, David Rothschildb (2017), Non-Representative Surveys: Modes, Dynamics, Party, and Likely Voter Space, url: https://researchdmr.com/
3. Sharad Goel, Adam Obeng, David Rothschild (2015), Non-Representative Surveys: Fast, Cheap, and Mostly Accurate, url: https://researchdmr.com/

❏ **Member Contributions**

Yingxin Zhang: EDAV; Data preprocessing; Implementation of models in R and Python (MRP, random forest, SGD Classifier, KNN, SVM, XGBoost); Coefficient plots visualization; Report composition.

Jiaying Zhang: EDAV; Model discussions; Reports compositions.

Jingyu Ren: EDAV and interpretation; used the beat model (SVM) to predict the new question; Reports compositions.

Qinwei Zhao: EDAV and interpretations; R code for coefficient plots; implementation of multiclass AUC score for model evaluation.

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.