To: Dr. NaLette Brodnax

From: Alex Adams

Subject: PPOL565 Memo 2: Analytical Plan

Date: April 16, 2021

1 Project Topic and Background

Over the past decade, healthcare has taken on a particular salience with regard to both politics and policy in the United States. One key component of this discourse centers on the role that the federal government (and, to a lesser extent, state and local governments), should play in providing access to quality affordable healthcare. On one hand, some activists on the political left believe that healthcare should be publicly subsidized for all Americans, effectively making it free (Diamond, 2021). In contrast, some on the right believe that not only should governments of all levels not fully subsidize healthcare, but that current government programs such as Medicare, Medicaid, or the Children's Health Insurance Program (CHIP) should be defunded or privatized (i.e. have the responsibility for their continued operation transferred away from government). The movement on the left has gained momentum in recent elections, with several freshmen and two-term members of the House of Representatives actively endorsing universal publicly-funded health care (most commonly under the banner of Medicare for All). Public support reached as high as 63% in the lead-up to the 2020 elections, creating increasing pressure for federal action to achieve universal health care (Jones, 2020). It is then an object of interest as to which factors are associated with or predict support for universal public health care, in order to more effectively deploy political capital and activist resources.

The intent of this project is to use a combination of parametric and non-parametric supervised machine learning techniques to predict a given respondent's preference for publicly subsidized healthcare relative to private healthcare using demographic, socioeconomic, and attitudinal features (see section 3: Methodology). This is a prediction task, but not a classification task, since the dependent variable (outlined in section 2: Data Sources) is continuous.

Some existing research suggests certain demographics may be more amenable to adopting a public health care system than others. A survey taken in Ireland found that women were more likely to support universal health care than men, and that those who are already enrolled in some level of government-provided healthcare are more likely to support it than their peers with private healthcare (Darker et al., 2018). More locally, research conducted in the wake of the passage of the Patient Protection and Affordable Care Act (commonly referred to as Obamacare) in 2010 found that African Americans were more likely to support universal health care than their white counterparts, and that this gap in opinion widened when president Obama adopted healthcare as a key issue for his administration. This effect is also present for other non-white racial groups, though the difference appears less pronounced. (Henderson & Hillygus, 2011; Tesler, 2012) Personal attitudes may also influence beliefs about whether healthcare should be publicly or privately funded. Jensen and Petersen observe a phenomenon they label the ”deservingness heuristic”, which describes how people see and evaluate others with regard to receiving social benefits such as unemployment payments or subsidized health care. They find that in general, sick recipients of healthcare-related social welfare benefits are perceived as more deserving than unemployed recipients of unemployment benefits. (Jensen & Petersen, 2017)

2 Data Sources

For this analysis, I will use data gathered by the RAND Corporation through the 2018 National Survey of Health Attitudes (NSHA), accessible through the Inter-University Consortium for Political and Social Research (ICPSR) hosted at the University of Michigan (Carman et al., 2019; RAND Corporation, 2021). This data set consists of 7,187 responses gathered from American adults in 2018. It includes attributes which encode demographic information (age, gender, race, education level, socioeconomic status), access to health care (i.e. whether or not a respondent has insurance, and if so, through whom and what type), and various attitudinal questions regarding healthcare and health outcomes. Many of the variables in this dataset are either categorical (typically ”yes/no” or ”strongly agree/agree/no opinion/disagree/strongly disagree”) or ordinal (ranking which factors seem most important, or whether something should be a minor or top-level priority).

My dependent variable of interest is $Q16{-}4$.[1] 1,651 respondents did not answer this question, leaving an effective data set of 5,536 responses. Relevant feature variables for this analysis have either been coded into ordinal brackets or are categorical. As such, I am only reporting descriptive statistics for two continuous feature variables (age and household size) in this table.

Table 1: Summary Statistics for $Q16 - 4$ and Two Relevant Feature Variables

Statistic	Q16-4	Age	Household Size
Count	5,536	5,536	5,529
Mean	39.767	54.972	2.577
Standard Deviation	27.401	15.664	1.464
Minimum	0	18	1
25th Percentile	17	43	2
50th Percentile	38	57	2
75th Percentile	56	67	3
Maximum	100	95	12

[1] To paraphrase: On a scale from 0 to 100, who should be responsible for providing health care, the federal government or private corporations? 0 indicates that the government should be completely responsible, 100 indicates the private sector should be solely responsible, and any values in between indicate a mix between the two.

The summary statistics in table 1 show that the median response to $Q16-4$ is less than 50, meaning that the majority of those surveyed believe the government should be more responsible for health care than private corporations. Figure 1 shows the distribution of the dependent variable:

Figure 1: Distribution of Responses to $Q16-4$

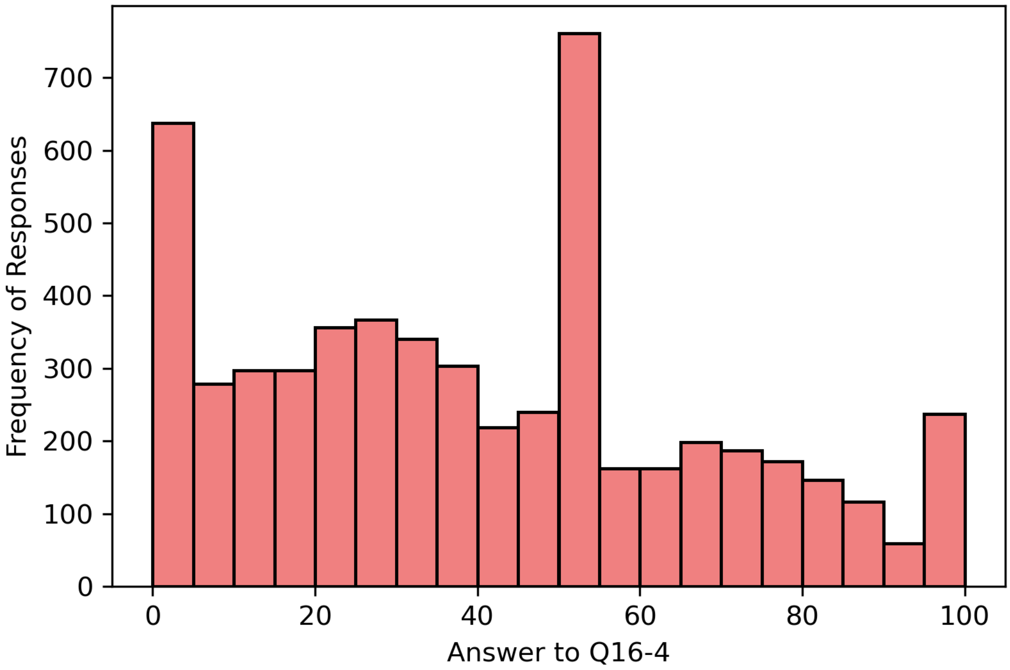


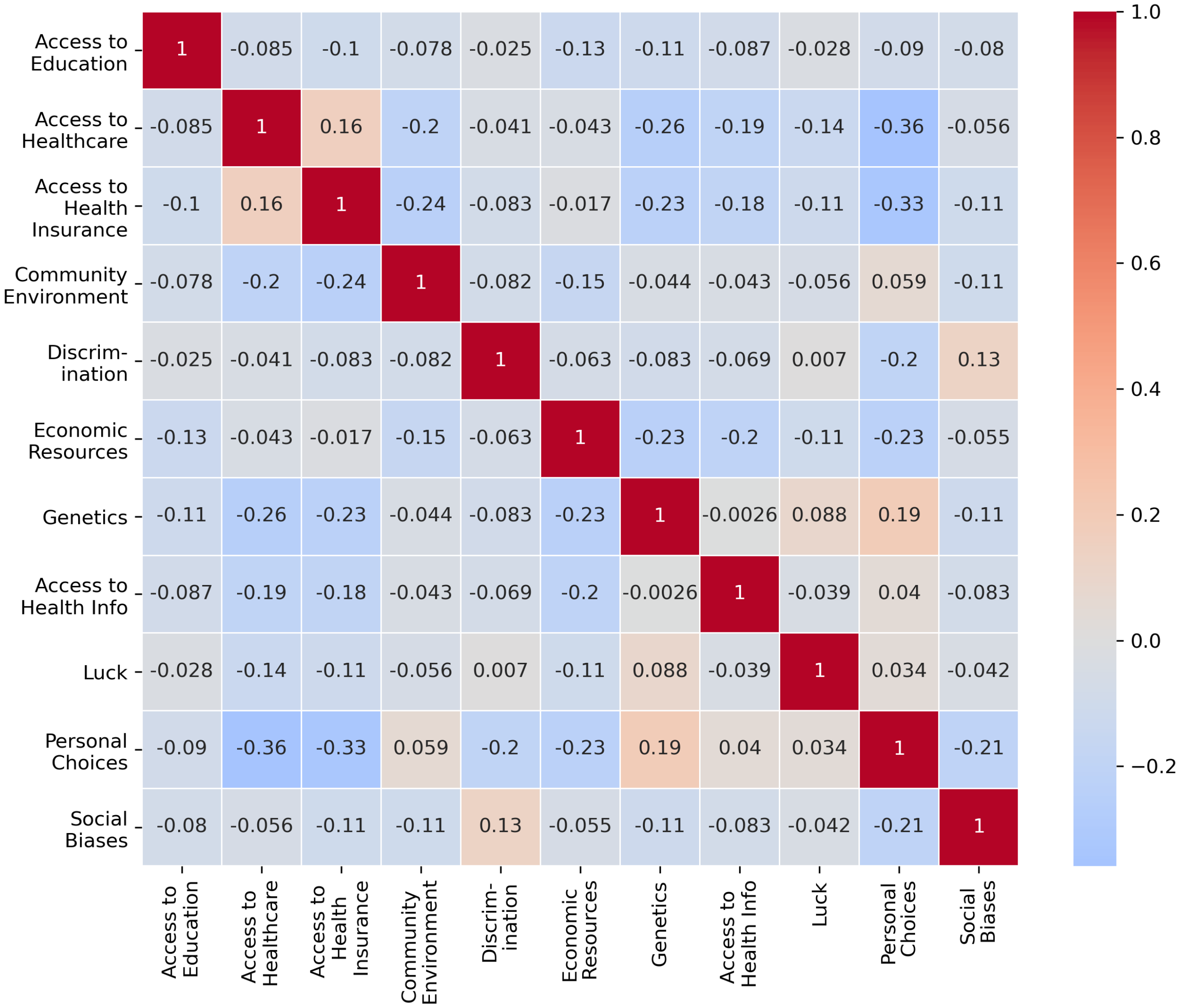
Figure 1: *This histogram shows the distribution of survey responses to* $Q16-4$*. 0 indicates a belief that the government should be solely responsible for health care, 100 indicates that private corporations should be held responsible, and all other values indicate a belief in shared responsibility.*

While many of the attributes included in this data set have small amounts of missing observations, the data set contains several thousand observations even when these are excluded or otherwise processed. Additionally, the machine learning methods I have selected for this analysis are robust against missing data, reducing the need to drop or impute certain values. In addition to standard demographic attributes (specifically race, gender, age, region of the U.S., education level, marital status, household size, family income, and employment status), my two primary independent variables of interest are items $Q11-B$[2] and $Q15-A$[3], which ask respondents which they weigh more heavily with regard to health outcomes, social factors or personal choices. Figure 2 displays a correlation heatmap between the different subvariables of $Q15-A$.

[2] Which of the following is the biggest reason that people in America become unhealthy? Personal Factors or External Factors?

[3] In the United States today, people with lower incomes live on average 7.5 years less than people with higher incomes. What do you think are the top three reasons why this is the case? (Select three options) Options: Access to a Good Education, Access to Healthcare, Access to Health Insurance, Community Environment, Discrimination, Economic Resources, Genetics, Access to Health Information, Luck, Personal Choices and Behavior, Treatment by Society of People with Low Incomes.

Figure 2: Q15-A Subvariable Correlations



	Access to Education	Access to Healthcare	Access to Health Insurance	Community Environment	Discrim-ination	Economic Resources	Genetics	Access to Health Info	Luck	Personal Choices	Social Biases
Access to Education	1	-0.085	-0.1	-0.078	-0.025	-0.13	-0.11	-0.087	-0.028	-0.09	-0.08
Access to Healthcare	-0.085	1	0.16	-0.2	-0.041	-0.043	-0.26	-0.19	-0.14	-0.36	-0.056
Access to Health Insurance	-0.1	0.16	1	-0.24	-0.083	-0.017	-0.23	-0.18	-0.11	-0.33	-0.11
Community Environment	-0.078	-0.2	-0.24	1	-0.082	-0.15	-0.044	-0.043	-0.056	0.059	-0.11
Discrim-ination	-0.025	-0.041	-0.083	-0.082	1	-0.063	-0.083	-0.069	0.007	-0.2	0.13
Economic Resources	-0.13	-0.043	-0.017	-0.15	-0.063	1	-0.23	-0.2	-0.11	-0.23	-0.055
Genetics	-0.11	-0.26	-0.23	-0.044	-0.083	-0.23	1	-0.0026	0.088	0.19	-0.11
Access to Health Info	-0.087	-0.19	-0.18	-0.043	-0.069	-0.2	-0.0026	1	-0.039	0.04	-0.083
Luck	-0.028	-0.14	-0.11	-0.056	0.007	-0.11	0.088	-0.039	1	0.034	-0.042
Personal Choices	-0.09	-0.36	-0.33	0.059	-0.2	-0.23	0.19	0.04	0.034	1	-0.21
Social Biases	-0.08	-0.056	-0.11	-0.11	0.13	-0.055	-0.11	-0.083	-0.042	-0.21	1

Figure 2: *This heatmap shows the correlations between possible factors survey respondents believe have the most influence on differential life expectancy between high-income and low-income Americans.*

Some of the correlations in Figure 1 follow logically. For example, choosing ”Personal Choices and Behavior” as one of the three factors which most explains income-based life expectancy gaps is inversely correlated with choosing ”Access to Healthcare” as another top three factor. The positive correlation between ”Discrimination” and ”Social Biases” also seems logical, as does the negative correlation between ”Personal Choices” and ”Social Biases”. What is less logical, however, is the positive correlation between ”Genetics” and ”Personal Choices”. Overall, none of these subvariable correlations are large, with the most extreme being between ”Access to Healthcare” and ”Personal Choices” at -0.36. This suggests that survey respondents may not have clear ideological attitudes regarding the personal-responsibility/social-welfare dichotomy, which could limit the predictive power of these subvariables for my dependent variable of interest.

I expect to find that respondents who fault structural outcomes more than personal choices will generally give lower scores to $Q16-4$, indicating they believe the federal government should take on greater responsibility for providing access to health care; I also expect the inverse to be true. The results for this analysis will depend in part on feature permutation importance; if demographic factors such as race or gender are the strongest predictors of support for public health care, then activists may find it relatively easy to reach out to members of those groups to drum up support. In contrast, if attitudinal or worldview based factors are most predictive, then the implications for activists and organizers are less clear.

3 Methodology

My research question is a supervised prediction task. To accomplish this task, I intend to use both logistic LASSO (Least Absolute Shrinkage and Selection Operator) regression, a decision tree algorithm, and a random forest algorithm from the *scikit-learn*(Pedregosa et al., 2011) package with a 70-30 training-test split and 8-fold cross-validation (8 is a factor of 5,536) to assess accuracy.

3.1 Parametric Technique: Logistic LASSO Regression

I considered both LASSO regression and ridge regression for this task. One of my feature variables of interest, $Q15-A$, is expressed in the data set as 11 separate variables. Each sub-variable is binary, and encodes whether the respondent selected the factor as one of the top three reasons they believe high-income people live longer than low-income people (with 1 indicating that the factor is in their top three, and 0 indicating that it is not). To these 11 sub-variables, I add nine demographic variables (several of which will be converted to a series of dummy variables) and one additional attitudinal variable, for approximately 45-50 attributes. I do not expect all of the feature variables in this analysis to have statistically significant associations with my dependent variable of interest. Unlike ridge regression, which preserves the total number of attributes in a data set, LASSO regression can apply a shrinkage penalty which reduces coefficient estimates to zero. This will reduce the variance of the model, make my results easier to interpret, and bolster any analysis of feature importance I may conduct during this project.

Signorino and Kirchner describe an example of LASSO implementation using data from the National Institutes of Health to predict whether or not a given individual will respond to a survey(Signorino & Kirchner, 2018). They find that LASSO regression closely mirrors the true distribution of survey response, regardless of whether the data is centered at the median values or not. They also observe that LASSO regression may not be effective at identifying important features from unimportant ones when the features exhibit high degrees of correlation. This reinforces my understanding that LASSO regression is a logical choice for this investigation, since Figure 1 demonstrates that many of my feature variables of interest are, at best, weakly correlated with each other. Perhaps most importantly, Signorino and Kirchner find that LASSO regression outperforms conventional linear regression on several metrics, particularly those related to classification (which, while not the focus of this analysis, could be a useful task for a follow-up inquiry).

3.2 Non-Parametric Technique: Decision Tree and Random Forest

For a non-parametric technique, I considered decision trees, K-nearest-neighbors, and random forest regressors. A decision tree is a machine learning algorithm which splits a data set into multiple subsets (or ”nodes”) based on different attributes, where each node corresponds to a particular region in p-dimensional space. The primary advantage of a decision tree, and why I have ultimately chosen to use it, is that its output is easily interpretable. While decision trees are susceptible to overfitting (when a model is fit to training data in a way which minimizes variance but restricts its predictive power for test data), they are highly conducive to clear and useful visualizations. Decision trees can also be tuned relatively easily, and can provide an approximation of feature importance.

Batterham and Christensen demonstrate the value of decision trees in analyzing factors which predict suicidal behavior among Australian adults (Batterham & Christensen, 2012). They grow different decision trees based on different waves within their broader population, and find that different factors predict suicidal ideation among different groups. It can be easily understood that for individuals with lower levels of depression, age is a significant predictor, while for individuals with higher levels of depression, histories of controlled substance use or abuse are more effective predictors. The ability to draw different conclusions about sub-populations within my data set could lead to a broader range of future lines of inquiry and potentially produce unexpected results.

I also intend to run a random forest regressor to supplement the decision tree model and obtain more accurate results. Random forests effectively ”grow” large numbers of decision trees using different sets of attributes, and then average the trees to arrive at an overall predictive model. While they are difficult to represent visually, random forests are more robust to overfitting and multicollinearity, and can produce more accurate predictions than an individual decision tree (Hanson et al., 2019). Since random forests take into account a large number of permutations of attributes, they can provide a more reliable order of variable importance than other machine learning algorithms.

References

Batterham, P. J., & Christensen, H. (2012). Longitudinal risk profiling for suicidal thoughts and behaviours in a community cohort using decision trees. *Journal of Affective Disorders*, *142*(1), 306–314. https://doi.org/https://doi.org/10.1016/j.jad.2012.05.021

Carman, K. G., Chandra, A., Weilant, S., Miller, C., & Tait, M. (2019). *2018 national survey of health attitudes: Description and top-line summary data*. RAND Corporation. https://doi.org/10.7249/RR2876

Darker, C. D., Donnelly-Swift, E., & Whiston, L. (2018). Demographic factors and attitudes that influence the support of the general public for the introduction of universal healthcare in ireland: A national survey. *Health Policy*, *122*(2), 147–156. https://doi.org/https://doi.org/10.1016/j.healthpol.2017.11.009

Diamond, D. (2021). House democrats bring back medicare-for-all, seeking to push biden left. https://www.washingtonpost.com/health/2021/03/16/house-democrats-medicare-for-all-biden/

Hanson, H. A., Martin, C., O'Neil, B., Leiser, C. L., Mayer, E. N., Smith, K. R., & Lowrance, W. T. (2019). The relative importance of race compared to health care and social factors in predicting prostate cancer mortality: A random forest approach. *Journal of Urology*, *202*(6), 1209–1216. https://doi.org/10.1097/JU.0000000000000416

Henderson, M., & Hillygus, D. S. (2011). The dynamics of health care opinion, 2008–2010: Partisanship, self-interest, and racial resentment. *Journal of Health Politics, Policy, and Law*, *36*(6), 945–960. https://doi.org/https://doi.org/10.1215/03616878-1460533

Jensen, C., & Petersen, M. B. (2017). The deservingness heuristic and the politics of health care. *American Journal of Political Science*, *61*(1), 68–83. https://doi.org/https://doi.org/10.1111/ajps.12251

Jones, B. (2020). Increasing share of americans favor a single government program to provide health care coverage. https://www.pewresearch.org/fact-tank/2020/09/29/increasing-share-of-americans-favor-a-single-government-program-to-provide-health-care-coverage/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

RAND Corporation. (2021, March 1). National survey of health attitudes, 2018. Retrieved March 1, 2021, from https://www.icpsr.umich.edu/web/ICPSR/studies/37633/summary

Signorino, C. S., & Kirchner, A. (2018). Using lasso to model interactions and nonlinearities in survey data. *Survey Practice*, *11*(1). https://doi.org/10.29115/SP-2018-0005

Tesler, M. (2012). The spillover of racialization into health care: How president obama polarized public opinion by racial attitudes and race. *American Journal of Political Science*, *56*(3), 690–704. https://doi.org/https://doi.org/10.1111/j.1540-5907.2011.00577.x