"KEEP YOUR GOVERNMENT HANDS OFF MY MEDICARE!": ANALYZING FACTORS WHICH PREDICT SUPPORT FOR PUBLICLY-FUNDED HEALTHCARE AMONG AMERICAN ADULTS

WHITE PAPER

Alexander Adams
McCourt School of Public Policy
Georgetown University
Washington, DC
aja149@georgetown.edu

May 14, 2021

ABSTRACT

At the core of the healthcare policy debate in the United States is the issue of public versus private funding of care. Many activists, especially on the American political left, are actively seeking to recruit politicians and candidates and convince voters to support their preferred position on healthcare. I trained a linear regression with an $L1$ penalty (also known as LASSO regression), a decision tree classifier, and a random forest classifier on data collected by the RAND Corporation in the 2018 National Survey of Health Attitudes ($n = 5,536$). I found that attitudinal factors (such as believing that personal factors explain health status and that personal choices are a significant factor in explaining income-related life expectancy disparities) are more strongly associated with opposition to publicly-funded healthcare than demographic factors or socioeconomic indicators like education level or employment status. I also found that elevating structural factors such as discrimination and access to health care in that same way is associated with support for publicly-funded healthcare. Additionally, being female or Hispanic or Latino is associated with higher support for government-funded healthcare, making these demographics potential targets for activist campaigns. However, none of the three modeling techniques used in this analysis produced highly accurate results, indicating that there may be other unobserved factors which influence attitudes toward healthcare.

Keywords Public Opinion · Healthcare · Decision Tree · Random Forest · LASSO Regression

1 Introduction

Over the past decade, healthcare has taken on a particular salience with regard to both politics and policy in the United States (Pew Research Center, 2021). 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) (Kliff & Nelson, 2017).

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 with aspects of classification and regression, 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). As such, it is possible that employment status or disability status could influence attitudes toward healthcare access.

2 Data Sources

For this analysis, I used 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] Indicate who you think should play a bigger role in providing healthcare services, the government or the private sector (businesses and nonprofits)? Responses: 0-100, 0 = The government should be solely responsible for providing health care, 100 = The private sector should be solely responsible for providing health care, Any value in between indicates a mixture of 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 $Q16\_4$

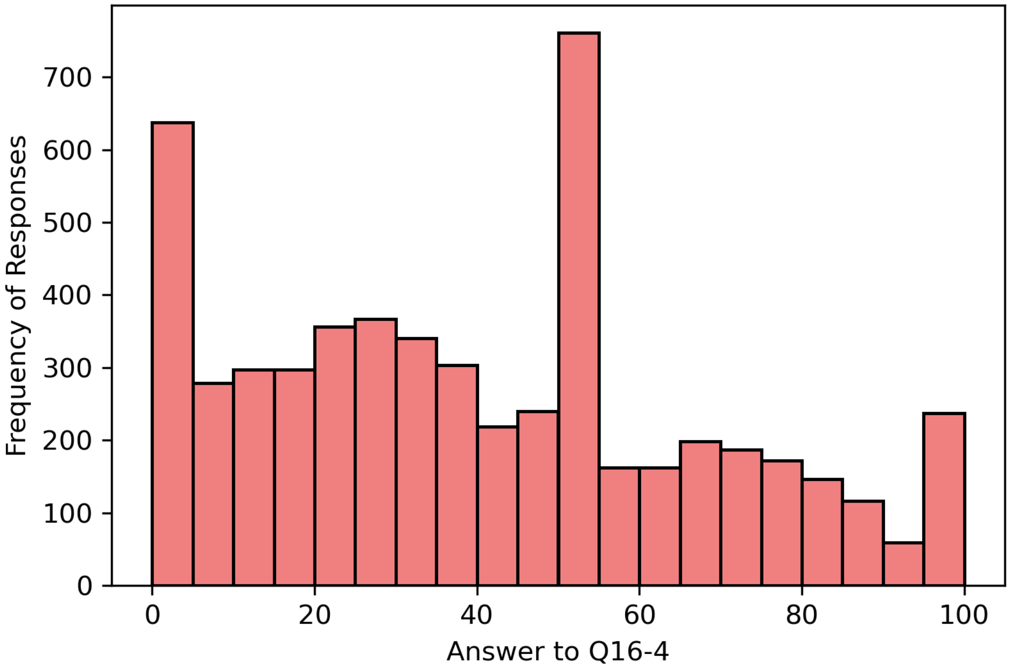


Figure 1: *This histogram shows the distribution of survey responses to $Q16\_4$. 0 indicates a belief that the government should be solely responsible for health care, 100 indicates that private corporations should be held responsible, and all other values indicate a belief in shared responsibility.*

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 (excepting the LASSO regression) 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 $Q11$[2] and $Q15A$[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 $Q15A$.

[2] Q11 consists of two subvariables, Q11A and Q11B. Q11A asks: Which statement is closer to your views? The biggest reason people in America become unhealthy is because they make poor choices that affect their health. OR The biggest reason people in America become unhealthy is because things outside of their control affect their health. Q11B asks: Which of the following is the biggest reason that people in America become unhealthy? Personal Factors or External Factors?

Due to the way responses are coded, and the similarity between the two question formulations, I combined these into a single feature variable.

[3] Q15A: 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: Q15A Subvariable Correlations

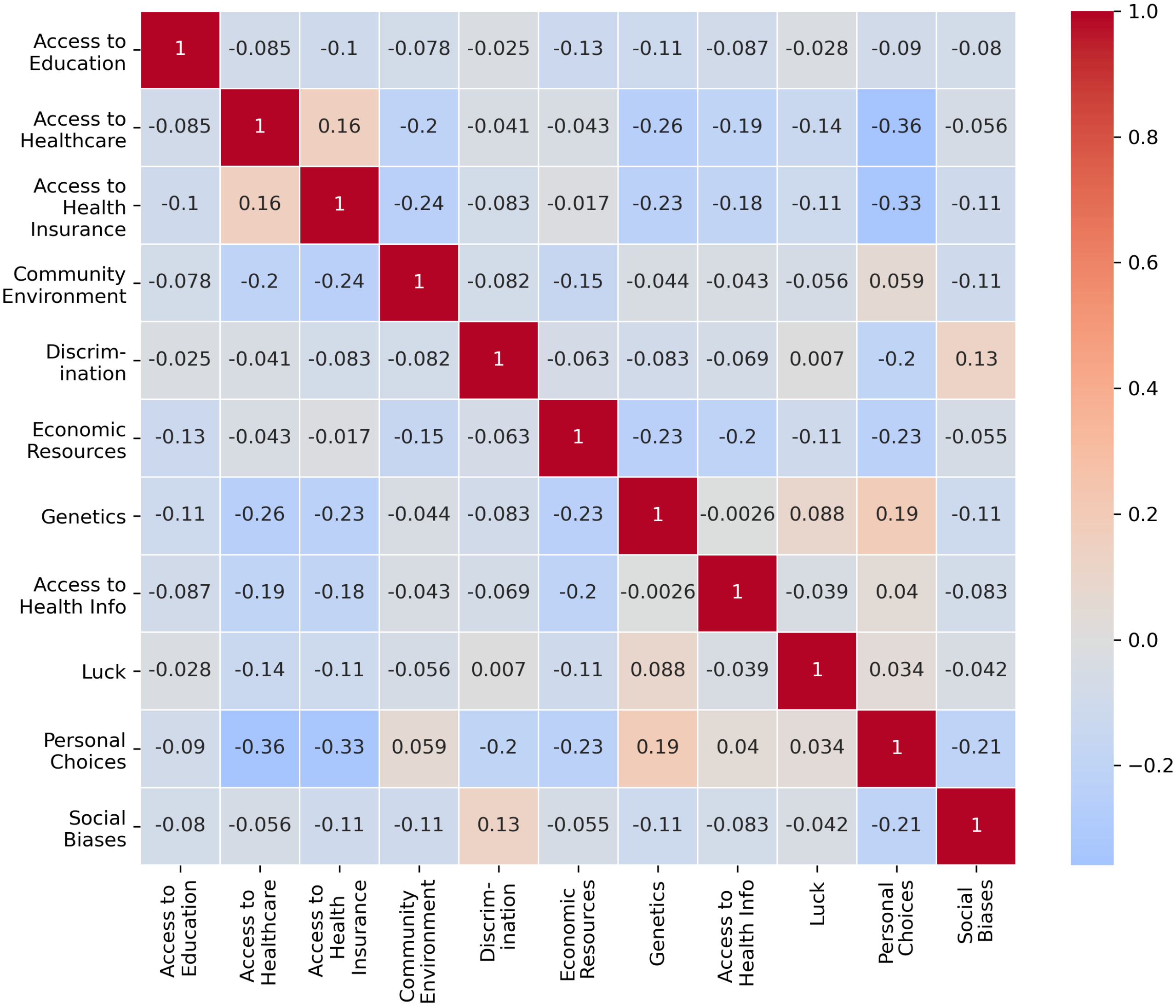


Figure 2: *This heatmap shows the correlations between possible factors survey respondents believe have the most influence on differential life expectancy between high-income and low-income Americans.*

Some of the correlations in Figure 2 follow logically. For example, choosing "Personal Choices and Behavior" (an inherently personal and individual-level factor) as one of the three factors which most explains income-based life expectancy gaps is inversely correlated with choosing "Access to Healthcare" (which is structural in nature) 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 or may have difficulty parsing out circumstances where individuals have agency from those where they do not. This 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 and classification 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 75-25 training-test split and 8-fold cross-validation (8 is a factor of 5,536, the number of observations I have after dropping those missing responses to the dependent variable) 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, $Q15A$, 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 a total of 38 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 2 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. I initially chose to use a decision tree regressor and a random forest regressor. However, the results those algorithms produced indicated that they performed significantly worse than a purely random guess. The decision tree in particular produced an R^2 of approximately -0.6, meaning that the residual sum of squares was greater than the true sum of squares, and the model was unable to generate useful predictions. To address this issue, I adapted this research design into a multi-class classification task. I recoded my dependent variable into three classes spanning the ranges 0-33, 34-66, and 67-100 (class labels 0, 1, and 2 respectively). I considered recoding based on percentiles, but as can be seen in figure 1, the distribution of this data shows biases in favor of 0, 50, and 100. This is unsurprising given the tendency of survey respondents to prefer round numbers or numbers which are multiples of 10 or 25 (Kettle & Häubl, 2010). Additionally, given the percentile ranks (the 33rd percentile response is 24, and the 66th percentile is 50), splitting into classes based on percentiles would result in one class which spans half the range of possible values, while the other two classes split the remainder.

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; indeed, I have included a full printout of the decision tree I used in this analysis in Appendix C. 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. I used validation curves to identify ideal values of two primary tuning parameters, maximum depth (the maximum number of times the tree can branch) and minimum samples per leaf (the minimum number of instances remaining in a given node to allow for splitting. I found that a maximum depth of 5 and a minimum number of samples per leaf of 80 produced the most accurate classification.[4]

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 ran a random forest classifier 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. For the random forest, I tuned the maximum depth hyperparameter to a depth of 7, and the number of estimators hyperparameter (how many trees to grow) to 600. I considered using the receiver operating characteristic (ROC) score to evaluate the decision tree and random forest. However, while calculating this score (and plotting the associated curve) is simple for a binary classification problem, it becomes more complex when more than two classes are involved. Ultimately I concluded that it would not provide me with information which was more useful than a cross-validated mean accuracy score, and so I did not calculate the ROC score for either nonparametric classifier.

4 Results

4.1 LASSO Regression

Because LASSO regression can result in some coefficients being adjusted to zero, it is useful for gaining understanding of feature importance and the directions of covariate relationships. The three variables with the most positive LASSO coefficients (see Appendix B) were $Q15A\_PCB$, $Q15A\_L$, and household income. Of these, $Q15A\_PCB$ is the only one to have a coefficient estimate above 1, and indeed produced an estimate greater than 3. This suggests that, *ceteris paribus*, choosing "Personal Choices and Behavior" as one of the three most significant factors in explaining income-related life expectancy disparities is associated with an increase of 3.316 in preference for privately-funded healthcare over publicly-funded healthcare.

The three variables which produced the most negative LASSO coefficient estimates were $Q15A\_AHC$, $Q15A\_AHI$, and $Q11$. $Q15A\_AHC$ resulted in the coefficient estimate with the greatest magnitude in this regression; choosing "Access to Healthcare" as one of the three most significant factors in explaining income-related life expectancy disparities is associated with a decrease of 4.537 in preference for privately-funded healthcare over publicly-funded healthcare. Being female, being hispanic, and choosing "Economic Resources", "Discrimination", or "Treatment by Society of Those with Low Incomes" as explaining income-related life expectancy disparities are also associated with coefficient estimates less than -1. With few exceptions, education level, employment status, marital status, and region of the country produced coefficient estimates of 0.

The LASSO regression produced an R^2 score of 0.203 for an α of 0.5. Testing with other values of α finds that R^2 approaches 0.209 as α goes to 0. This value of R^2 is particularly low, and indicates that the feature variables selected for this analysis do not have strong predictive or explanatory power. It also suggests that LASSO regression may not have been significantly more effective than a standard ordinary least squares (OLS) linear regression at modeling the relationship between the feature variables of interest and the outcome variable in this analysis.

4.2 Decision Tree

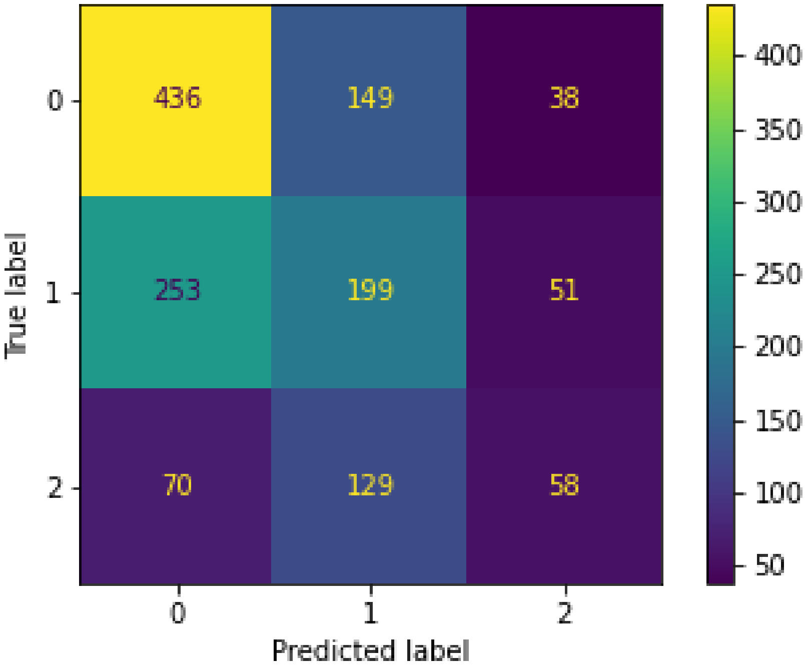
While classification is a different task from regression, the two tasks are comparable in one way: feature importances. Like the LASSO regression, the decision tree classifier found that $Q15A\_PCB$, $Q15A\_AHC$, household income, and $Q11$ had high feature importances, indicating that the model relied heavily on these attributes to classify instances. $Q15A\_PCB$ is the feature variable which forms the first split in the decision tree. The decision tree also evaluated age as having the second-highest feature importance of any attribute in this data set. Furthermore, the decision tree

[4] Note: For both the Decision Tree and the Random Forest classifiers, random state = 10

found that marital status, education level, employment status, region of the United States, and most racial categories had feature importance scores of 0, indicating that the model did not rely on them at all when classifying instances.

The decision tree produced a mean accuracy score of 0.509. This output means that the decision tree was able to correctly classify a given instance as having a low (0-33), medium (34-66), or high (67-100) score on healthcare preference only about half the time. This outcome is also reflected in the gini indices of different nodes throughout the tree. For most nodes, the gini index of the node is somewhere between 0.4 and 0.6, with the maximum gini index being 0.666 and the minimum being 0.296. Every node in the tree contains at least one instance of all three possible classes. Figure 3 below shows the confusion matrix for the decision tree:

Figure 3: Decision Tree Confusion Matrix



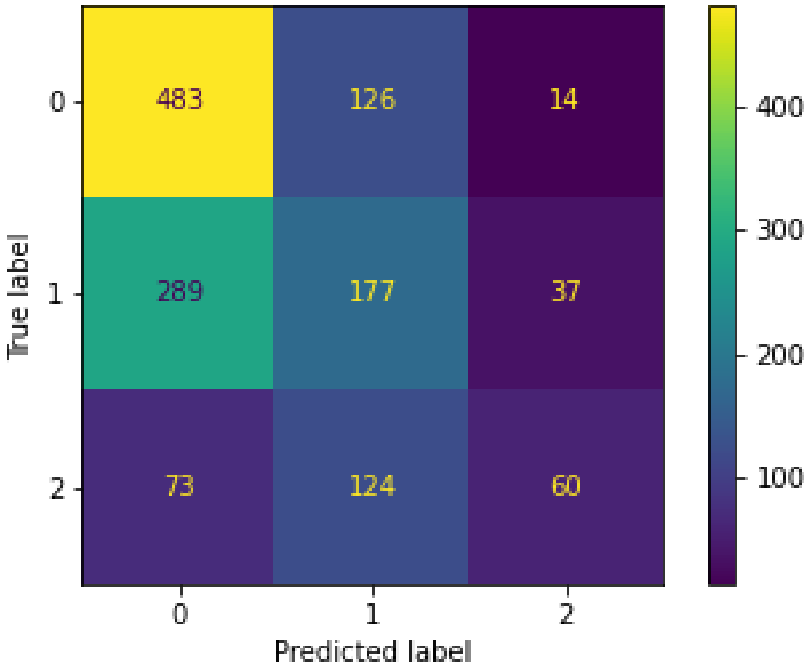
Of particular note here is the bottom row, which shows the predicted labels for instances of class 2. For this class, the correct label was the least likely to be predicted, and occurred half as often as a prediction of 1 for a class 2 instance. This suggests that the decision tree experienced difficulty in differentiating between a medium or high score on the dependent variable. The decision tree also predicted a label of 0 more than a label of 1 for class 1 instances. Overall, this confusion matrix indicates that the decision tree was able to correctly classify instances of class 0 (the modal class) with a reasonable degree of accuracy, but was unable to effectively classify instances from the other two (less prevalent) classes.

4.3 Random Forest

The random forest classifier produced a mean accuracy score of 0.516. This score is almost identical to that of the decision tree, meaning that the random forest did not meaningfully outperform the decision tree on this task. Similarly to the decision tree, the random forest classifier identified $Q15A\_PCB$, age, $Q15A\_AHC$, $Q15A\_AHI$, household income, and $Q11$ as being the most important features in executing the classification task. Given the number of trees grown by the random forest classifier (600), it is unsurprising that every feature variable in this analysis has a non-zero feature importance score (in contrast to the decision tree, where the majority of features had an importance score of zero).

Given the similarity in performance between the decision tree and the random forest, it makes sense to evaluate the classification results to determine if the two models made similar mistakes. Figure 4 shows the confusion matrix output for the random forest classifier.

Figure 4: Random Forest Confusion Matrix



Like the decision tree, the random forest classifier was most successful at classifying instances of class 0. However, while the random forest was more successful at classifying class 0 and class 2 instances than the decision tree, it was actually *less* successful at classifying class 1 instances. The decision tree correctly classified 22 more class 1 instances than the random forest, while the random forest only correctly classified 2 more class 2 instances than the decision tree. Additionally, by calling the *predict* attribute, I was able to generate a list of predicted class labels for the decision tree and random forest. Out of 1,383 instances in the test data, the two classifiers predicted the same label for 1,103 instances[5], for an inter-model agreement rate of 79.75%. I also identified the instances where both the decision tree and the random forest predicted the same class label, and that class label was correct. The two models had an inter-model agreement and accuracy rate of 43.02%. This suggests that roughly 1 in 5 instances correctly classified by the decision tree were incorrectly classified by the random forest, and vice versa.

5 Conclusions

The intent of this project was to compare attitudinal, demographic, and socioeconomic indicators to predict support for publicly-funded healthcare. Based on the results from the LASSO regression, the decision tree classifier, and the random forest classifier, attitudinal factors (specifically those related to beliefs about personal versus structural factors) are the strongest predictors and indicators of support for greater government involvement in health care. However, not all attitudinal factors are created equally: of the 11 subvariables involved in $Q15A$, only a few ("Personal Choices and Behavior", "Access to Health Care", and "Access to Health Information") were consistently relevant across all three model specifications. The other primary attitudinal variable in my feature set, $Q11$, was also relevant, indicating that a preference for structural factors over personal factors (or vice versa) informs attitudes toward which entities should fund health care.

The specific results from the LASSO regression comport with prior expectations regarding healthcare policy stances. Common rhetoric in left-leaning policy spaces grants that structural factors are more significant than personal ones. This matches the important $Q15A$ subvariables; in the United States, which has a high level of income inequality and no universal healthcare system (and where the costs of private insurance can be prohibitively high), access to

[5] Note: This does not mean that either model predicted the correct label for a given instance, just that the two models each predicted the same label for that instance.

healthcare (and health information) is a structural factor, rather than a personal one. Conversely, those on the political right (who often favor solutions from the private sector and prefer to minimize government involvement) tend to prioritize individual-level characteristics, like personal choices and behavior. The results bear this out: viewing access to healthcare as important is associated with lower $Q16\_4$ scores, indicating greater favorability for government-provided healthcare, while viewing personal choices and behavior as important is associated with higher $Q16\_4$ scores, indicating greater favorability for privately-funded healthcare.

Among the demographic variables included in the set of feature variables used in this investigation, gender and Hispanic racial identity were the most consistently relevant across different model forms. Female respondents and respondents who self-identified as Hispanic or Latino preferred that the government take on a greater role in providing healthcare than male respondents and those who did not self-identify as Hispanic or Latino. Household income was also somewhat significant, though in the opposite direction. The nature of this variable as an ordinal item with uneven gradiations means that it is difficult to determine an exact coefficient or quantifiable relationship between household income and support for publicly-funded healthcare. What can be concluded, though, is that as income increases, support for government involvement in healthcare decreases.

The primary question of this project concerns the practical implications of these findings. Voter outreach based on attitudinal attributes is impossible without other signifiers, such as organizational membership or charitable donations. Theoretically, an group of advocates or activists could identify social institutions associated with greater credibility in the impact of structural factors and target their members with informational materials, but such voter outreach would likely be limited to political actors with high levels of resources and robust community networks. More broadly useful are the results regarding demographic trends. Based on the model outputs, activists seeking to advance the cause of government-provided healthcare should target women and Hispanic or Latino Americans. Fortunately for proponents of government-funded healthcare, those are two groups which are increasingly relevant in the American electorate. The literature suggests that as more women achieve political power and name recognition, the number of female candidates for office increases, as does women's political participation (Burns et al., 2018; Jensen & Petersen, 2017). The candidacy of Hillary Clinton in 2016 was a particularly strong motivator for many women in the United States (Bonneau & Kanthak, 2020). The Latino population in the United States is consistently one of the fastest-growing segments of the nation, and has been for over two decades (Chapa & Rosa, 2004; de Haymes & Kilty, 2007). In the past, this population growth was largely due to immigration; while that still plays a factor, recent trends show that an increasingly large number of Latinos in the United States are native-born. This trend is likely to continue, meaning that Latino Americans could be a rich target for engaging in political activism around the issue of healthcare (Krogstad & Lopez, 2020). The third demographic variable which was most strongly associated with a definitive impact on preference for or against publicly-funded healthcare was household income, with higher-income respondents preferring private-sector control and lower-income respondents preferring public-sector control. Activists could reach out to low-income communities and low-wage workplaces. However, given the interplay between capital and political power in the United States, this may be less likely to result in active engagement with the cause of universal public health care than pursuing other groups with more resources. This is not to say that lower-income individuals do not have political agency, but rather that activists have limited time and energy, and should invest those things into the people and groups which are most likely to be able to contribute meaningfully toward their cause.

None of the regression or classification algorithms used in this analysis were particularly effective. While each provides useful insights into the relationships between different feature variables, the relatively low accuracy rates and R^2 scores lead me to conclude that there are unobserved variables (either present in the raw data but not selected as a feature variable for this analysis or not recorded during the data collection process) which explain significant amounts of variation in the dependent variable. As such, this analysis does not lend itself to absolute or unambiguous recommendations in the policy or activism spheres, and this question would benefit from further study.

There are two fundamental limitations of this analysis. The first is that the data used is from 2018. With the advent of technology and the 24-hour news cycle, public opinion can shift rapidly. Thus, while the trends present in this data set reflect public opinion in 2018, it is possible that the evolving discourses around health care in the United States (particularly since the onset of the COVID-19 pandemic in early 2020) have shifted public views, decreasing the relevance of these conclusions.

The second limiation is perhaps even more significant. The original survey from which feature variables were selected does not include an item which records partisanship or political affiliation. The increasing nationalization of American politics means that even mundane issues are frequently viewed through a partisan lens, with partisans adopting the position of their preferred party. (Sievert & McKee, 2019). This nationalization of politics mirrors a parallel trend of increasing polarization developed through negative partisanship (i.e., defining one's partisan stances through contrast with an out-group rather than positive identification with an in-group).(Abramowitz & Webster, 2016; Caughey et al., 2018). Regarding the most recent high-profile effort by the federal government to expand healthcare access (the Patient

Protection and Affordable Care Act), self-identified Democrats are more likely to credit the law for its positive effects, and self-identified Republicans are more likely to criticize the law for its shortcomings. (McCabe, 2016). In one study, 3 in 4 Democrats prefer a health care system run by the federal government, while 4 in 5 Republicans oppose such a system and favor private sector involvement. (Blendon et al., 2021). Partisanship clearly informs public opinion, particularly on issues as salient as health care. The lack of an indicator of partisanship in this data set is a significant weakness of this analysis, and more generally of the NSHA as an instrument.

Future work on this topic should include indicators of political alignment or partisan affiliation (either "liberal" versus "conservative" or "Democrat" versus "Republican"), as well as specific questions around particular instances of legislation (such as the Patient Protection and Affordable Care Act or the 2017 American Health Care Act). Until such surveys or polls are conducted, pro-universal-public-healthcare activists would do well to focus their efforts on women, Latinos, and those with low-incomes.

6 Acknowledgements

The author would like to thank Dr. NaLette Brodnax, Assistant Professor at the McCourt School of Public Policy at Georgetown University. The author would also like to thank the program directors of the Master of Science in Data Science for Public Policy program at the McCourt School, as well as the faculty, staff, and administrators of the McCourt School.

7 Appendices

7.1 Appendix A: Codebook

4YearDegree: Respondent’s highest level of education is a four-year degree or higher

Age: Age in years

AAPI: Respondent is Asian American or Pacific Islander

Assoc_Degree: Respondent’s highest level of education is some college, including an associate’s degree

Black: Respondent is Black or African American

Disabled: Respondent is disabled

Divorced: Respondent is divorced

Female: Respondent is female

HHoldSize: Respondent’s household size

HHIncome: Respondent’s annual family income. 8-level ordinal variable. For specific ranges, see 37633-0001-Codebook-ICPSR, ’DEM_FAMILYINCOME_SHORT’

HighSchool: Respondent’s highest level of education is a high school diploma

Hispanic: Respondent is Hispanic or Latino

Indigenous: Respondent is Native American or other Indigenous American

NeverMarried: Respondent has never been married

Midwest: Respondent is from the Midwest region of the United States

Other_Nonworking: Respondent was not employed at time of data collection for unspecified reasons

Other_Races: Respondent self-identifies as racial category besides White, Black, Hispanic, Asian, or Native

American

Q11: Which set of factors is more responsible for Americans becoming unhealthy? 0 = Personal Factors, 1 = Environmental/Societal Factors

Q15A_AHC: Respondent selected "Access to Health Care" as one of the top three explanations for disparities in life expectancy between high-income and low-income people

Q15A_AHI: Respondent selected "Access to Health Information" as one of the top three explanations for disparities in life expectancy between high-income and low-income people

Q15A_ATGE: Respondent selected "Access to a Good Education" as one of the top three explanations for disparities in life expectancy between high-income and low-income people

Q15A_CE: Respondent selected "Community Environment" as one of the top three explanations for disparities in life expectancy between high-income and low-income people

Q15A_D: Respondent selected "Discrimination" as one of the top three explanations for disparities in life expectancy between high-income and low-income people

Q15A_ER: Respondent selected "Economic Resources" as one of the top three explanations for disparities in life expectancy between high-income and low-income people

Q15A_GBM: Respondent selected "Genetics (someone’s biological makeup)" as one of the top three explanations for disparities in life expectancy between high-income and low-income people

Q15A_HITH: Respondent selected "Health Information They Have" as one of the top three explanations for disparities in life expectancy between high-income and low-income people

Q15A_L: Respondent selected "Luck" as one of the top three explanations for disparities in life expectancy between high-income and low-income people

Q15A_PCB: Respondent selected "Personal Choices and Behavior" as one of the top three explanations for disparities in life expectancy between high-income and low-income people

Q15A_TBSTLI: Respondent selected "Treatment by Society of Those with Low Incomes" as one of the top three explanations for disparities in life expectancy between high-income and low-income people

Retired: Respondent is retired

R_NEast: Respondent is from the northeastern United States

SelfEmployed: Respondent is self-employed

Separated: Respondent is separated from their spouse, but is not formally divorced

South: Respondent is from the southern United States

TempLayoff: Respondent is temporarily laid off from work

Unemployed: Respondent is unemployed

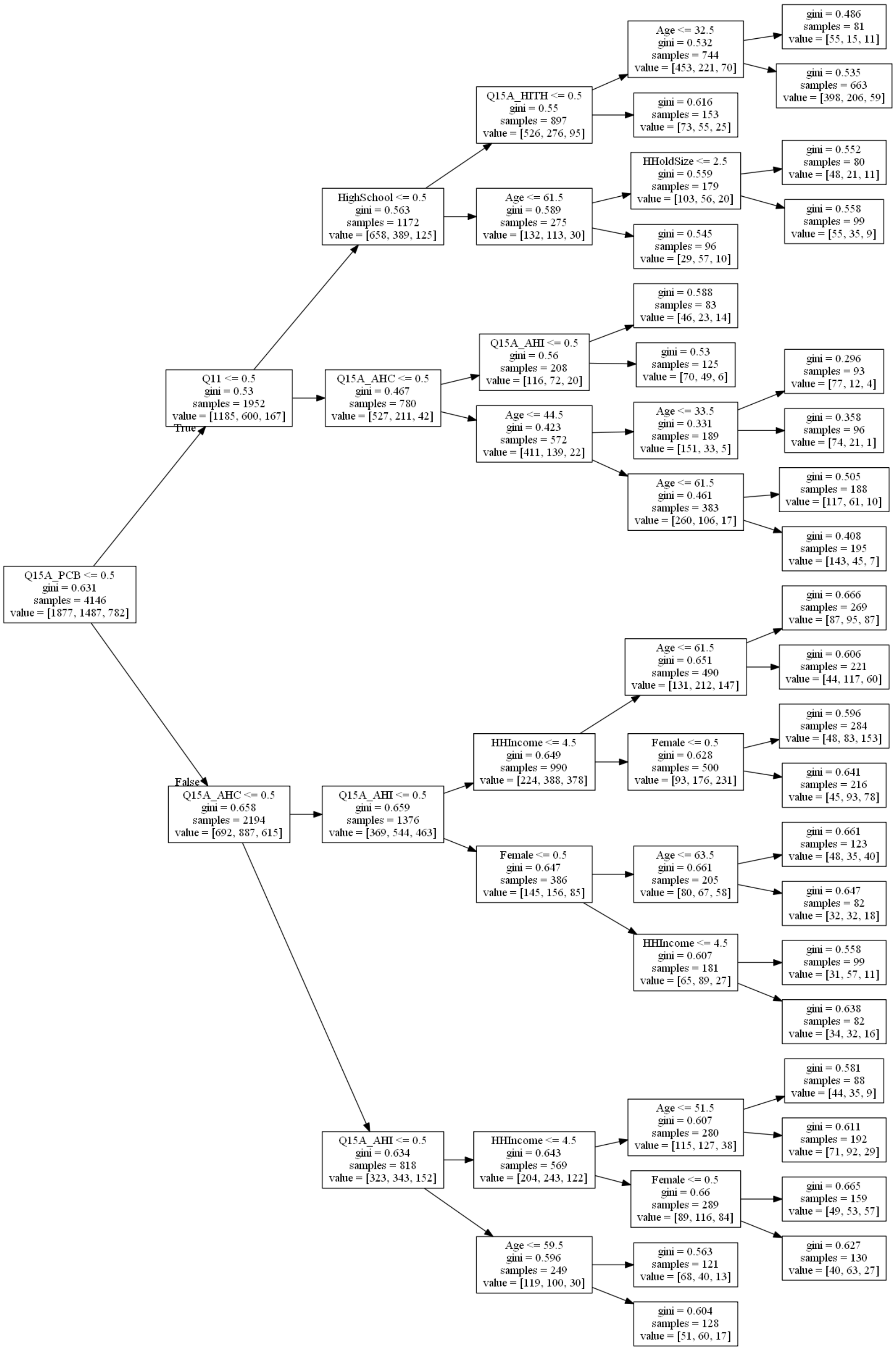
West: Respondent is from the western United States

Widowed: Respondent is widowed

7.2 Appendix B: Full LASSO Regression Coefficient Output

Variable	LASSO Coefficient
Q15A_PCB	3.316622
Q15A_L	0.923623
HHIncome	0.879599
Other_Nonworking	0.714098
Q15A_GBM	0.448295
HHoldSize	0.406971
Q15A_HITH	0.263953
Age	0.147193
Other_Races	0.096138
Separated	-0.000000
HighSchool	0.000000
Indigenous	0.000000
AAPI	0.000000
Disabled	-0.000000
Retired	0.000000
TempLayoff	-0.000000
SelfEmployed	0.000000
Widowed	-0.000000
4YearDegree	-0.000000
Assoc_Degree	-0.000000
West	0.000000
Q15A_CE	0.000000
Midwest	-0.000000
South	0.000000
NeverMarried	-0.051157
Unemployed	-0.075867
Divorced	-0.245887
Black	-0.458220
R_NEast	-0.461751
Q15A_ATGE	-0.654189
Female	-1.295236
Hispanic	-1.325997
Q15A_ER	-1.337334
Q15A_D	-1.535579
Q15A_TBSTLI	-1.775627
Q11	-2.868413
Q15A_AHI	-3.512656
Q15A_AHC	-4.537248

7.3 Appendix C: Decision Tree



7.4 Appendix D: Feature Importances

Variable	Decision Tree	Random Forest
Q15A_PCB	0.526350	0.164789
Age	0.110437	0.093577
Q15A_AHC	0.101039	0.090221
Q15A_AHI	0.073083	0.067186
HHIncome	0.066846	0.062629
Female	0.054257	0.032208
Q11	0.035333	0.074056
HighSchool	0.017809	0.015513
Q15A_HITH	0.012728	0.022163
HHoldSize	0.002117	0.039224
Black	0.000000	0.015322
Other_Nonworking	0.000000	0.010007
Divorced	0.000000	0.009251
Disabled	0.000000	0.010616
Hispanic	0.000000	0.018390
Retired	0.000000	0.013713
AAPI	0.000000	0.007363
Unemployed	0.000000	0.006411
TempLayoff	0.000000	0.001475
SelfEmployed	0.000000	0.009882
Indigenous	0.000000	0.002872
NeverMarried	0.000000	0.010658
Widowed	0.000000	0.010631
West	0.000000	0.011361
Separated	0.000000	0.004487
4YearDegree	0.000000	0.017126
Assoc_Degree	0.000000	0.010319
South	0.000000	0.011303
Midwest	0.000000	0.010682
R_NEast	0.000000	0.011181
Q15A_TBSTLI	0.000000	0.019132
Q15A_L	0.000000	0.013340
Q15A_GBM	0.000000	0.030676
Q15A_ER	0.000000	0.018911
Q15A_D	0.000000	0.019439
Q15A_CE	0.000000	0.017600
Q15A_ATGE	0.000000	0.010495
Other_Races	0.000000	0.005789

References

Abramowitz, A. I., & Webster, S. (2016). The rise of negative partisanship and the nationalization of u.s. elections in the 21st century. *Electoral Studies*, *41*, 12–22. https://doi.org/https://doi.org/10.1016/j.electstud.2015.11.001

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

Blendon, R. J., Benson, J. M., & Schneider, E. C. (2021). The future of health policy in a partisan united states: Insights from public opinion polls. *JAMA*, *325*(13), 1253–1254.

Bonneau, C. W., & Kanthak, K. (2020). Stronger together: Political ambition and the presentation of women running for office. *Politics, Groups, and Identities*, *8*(3), 576–594. https://doi.org/10.1080/21565503.2018.1528159

Burns, N., Schlozman, K., Jardina, A., Shames, S., & Verba, S. (2018). What's happened to the gender gap in political participation?: How might we explain it? *100 years of the nineteenth amendment* (pp. 69–104). Oxford University Press. https://doi.org/10.1093/oso/9780190265144.003.0004

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

Caughey, D., Dunham, J., & Warshaw, C. (2018). The ideological nationalization of partisan subconstituencies in the american states. *Public Choice*, *176*(1), 133–151.

Chapa, J., & Rosa, B. D. L. (2004). Latino population growth, socioeconomic and demographic characteristics, and implications for educational attainment. *Education and Urban Society*, *36*(2), 130–149. https://doi.org/10.1177/0013124503261320

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

de Haymes, M. V., & Kilty, K. M. (2007). Latino population growth, characteristics, and settlement trends: Implications for social work education in a dynamic political climate. *Journal of Social Work Education*, *43*(1), 101–116. https://doi.org/10.5175/JSWE.2007.200400493

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/

Kettle, K., & Häubl, G. (2010). Numeric fluency and preference. *ACR North American Advances*.

Kliff, S., & Nelson, L. (2017). The american health care act: The republicans' bill to replace obamacare,nbsp;explained. https://www.vox.com/2017/3/6/14829526/american-health-care-act-gop-replacement

Krogstad, J. M., & Lopez, M. H. (2020). Hispanic nativity shift. https://www.pewresearch.org/hispanic/2014/04/29/hispanic-nativity-shift/

McCabe, K. T. (2016). Attitude responsiveness and partisan bias: Direct experience with the affordable care act. *Political Behavior*, *38*(4), 861–882.

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.

Pew Research Center. (2021). Important issues in the 2020 election. https://www.pewresearch.org/politics/2020/08/13/important-issues-in-the-2020-election/

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

Sievert, J., & McKee, S. C. (2019). Nationalization in us senate and gubernatorial elections. *American Politics Research*, *47*(5), 1055–1080.

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