Predicting County-Level COVID-19 Vaccine Hesitancy

Prayag Chatha, Josh Jiang, Declan McNamara

December 05, 2021

1   Introduction

Vaccine hesitancy is a major impediment to the public health response to COVID-19. Mistrust of the COVID-19 vaccine can originate from many complex sources that can be political, historical, racial, etc. In this study, we examine the relationship between vaccination rates in US counties and demographic data. By performing a thorough quantitative study of this phenomenon, we hope to identify key factors that lead to vaccine hesitancy. Identifying these factors can be useful in crafting public health messaging to promote vaccination.

2   Data

2.1   Data Sources

We derived our data on county-level vaccination rates[1] from a scrape of the CDC's county view dashboard dating back to June 19th, 2021. We chose this date as roughly half of the US population had been fully vaccinated by this point in time.[2] For potential predictors of county-level vaccination, we obtained county-level results for the 2020 US Presidential election[3], demographic data (e.g. racial composition, income levels) from the 2015 5-year Census estimate[4], as well as county education levels (i.e. prevelance of a bachelor's degree or higher).[5]

2.2   Data Overview

Our combined data set contains complete data on 2,809 counties[6] in 47 states (sans Alaska, Hawaii, and Texas) as well as Washington, DC. These counties ranged from metropolitan areas[7] such as Los Angeles, Cook, and Maricopa counties to rural areas containing only a few hundred inhabitants. As counties primarily represent geographic area, most counties tend to be less-populous than average.

These counties had a mean and median adult vaccination rate of 41%, with most counties falling within a range of about 30-50% vaccination. We inspected several counties with extremely high or low vaccination levels.

---

[1] https://github.com/covidestim/cdc-vaccine-data
[2] https://ourworldindata.org/covid-vaccinations?country=USA
[3] https://www.kaggle.com/etsc9287/2020-general-election-polls
[4] https://www.kaggle.com/muonneutrino/us-census-demographic-data
[5] https://www.ers.usda.gov/data-products/county-level-data-sets/
[6] Out of the approximately 3,100 counties and equivalent municipalities in the USA
[7] In New York City, vaccination data for all five boroughs were combined, which led to its exclusion from the final data set.
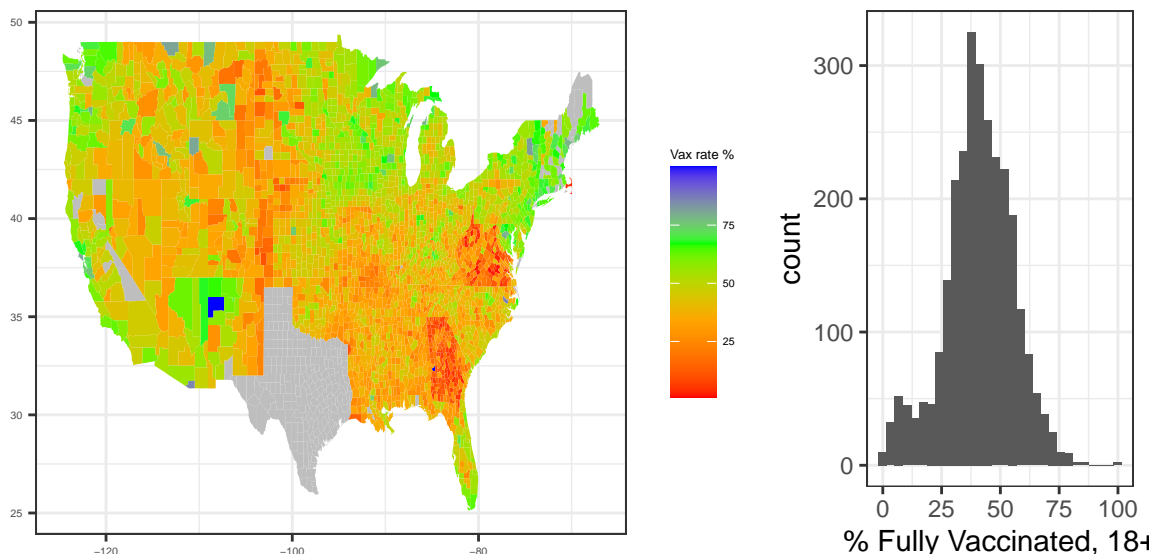
Figure 1: Map and Distribution of 18-Plus vaccination Rates by County, June 19th, 2021

2.2.1 Counties with Unusually High Vaccination Rates

The four counties with the highest vaccination rates–McKinley NM (100%), Chattahoochee GA (100%), Martin NC (86%), and Santa Cruz AZ (85%)–are all rural areas with large minority (Native American, Black, or Hispanic) populations. Several highly affluent counties, such as Montgomery MD, Marin CA, and Glacier MT had a vaccination rate of 80% or above. This suggests that counties that were particularly hard-hit by Covid-19 (i.e., poor and less-white areas) saw widespread vaccination adoption, but also that wealth correlates with vaccination. The counties with the highest vaccination rates generally leaned Democratic in the 2020 presidential election.

2.2.2 Counties with Unusually Low Vaccination Rates

The counties with the lowest vaccination rates (0-2%) were mostly in rural (i.e. western) Virginia, such as Appomattox County (1.7%) and the city of Lynchburg (0.8%). Other counties at the bottom end of the range included the island of Nantucket, MA (1.2%) and Morgan County, WV (2.1%). Remarkably, the 50 counties with the lowest vaccination rate (6% or lower) are all located in either Virginia, Georgia, West Virginia, or the Cape Cod region of Massachusetts. These include both mostly-white and white-minority counties, though most of them leaned Republican in the 2020 presidential election.

2.3 Exploratory Data Analysis

To dig further into demographic patterns of vaccine hesitancy, we figure 2 shows a scatterplot matrix for features roughly corresponding to education, ethnicity, wealth, political tendency, and population density. Blue points indicate counties with an above-average ($> 41\%$) vaccination rate, while counties with a below-average vaccination rate are colored red. Income per capita and the percentage of adults with a college education appear to predict a high vaccination rate, while strong support for Donald Trump in the 2020 presidential election correlates negatively with vaccination. The counties with the largest populations (i.e., cities) tend to have above-average vaccination rates, though this relationship is somewhat weak. Somewhat surprisingly, a county's percentage of white population seems to be non-predictive of vaccine adoption. It seems plausible that a model based only on education, income, or the 2020 election results could do a decent job of classifying counties.

To further investigate patterns in vaccine adoption, we compared adult vaccination rates against a variety of demographic features. Figure 2 shows six of the more interesting correlation scatterplots. We see that

an educated and high-earning populace tends to predict a high vaccination rate, both having a correlation coefficient of $\rho = 0.41$. In contrast, the more a county supported Donald Trump in the 2020 Presidential Election, the lower vaccination rates tend to be ($\rho = -0.38$). We speculate that a college education would be associated with a trust of scientific and media institutions (and thus amenability to vaccination). Furthermore, vaccination may be less accessible to poorer Americans, who might not be able to take time off of work or fear hidden medical costs. It is unsurprising that political leanings correlate so strongly with vaccination rates given that attitudes towards public health measures (e.g. mask wearing, social distancing) have been politicized since nearly the start of the pandemic.

Of the several variables pertaining to a county's racial composition, the percent of Black population had the strongest relationship with vaccination rates, showing a somewhat negative correlation ($\rho = -0.25$.). Similarly, the percentage of the workforce engaged in production (i.e., industrial or agricultural work) was a negative predictor ($\rho = -0.20$.) These suggest some disparities in either access to (or trust of) inoculation efforts.[8] We also observed that (log-scaled) population was associated with vaccination ($\rho = 0.33$), suggesting more vaccination in urban areas. More populated counties (i.e. urban Ameria) tended to vote for Joe Biden in the last election, so this effect could be merely a reflection of existing political polarization.
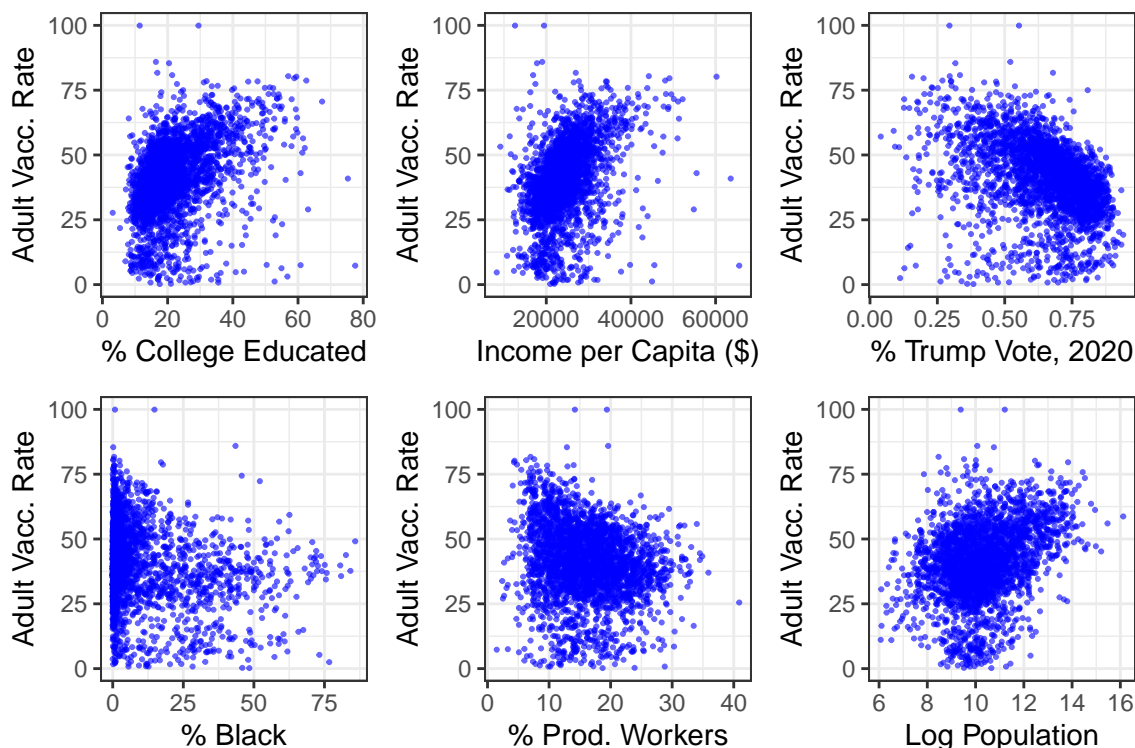

Figure 2: Correlation Scatterplots for Select Demographic Features

3 Prediction

3.1 AdaBoost

AdaBoost, short for "adaptive boosting," is the classical[9] boosting algorithm for classification. Following the intuition that an ensemble of weak learners can act in concert as a strong learners, at each successive iteration AdaBoost assigns additional loss weight to previously-misclassified observations. In theory, AdaBoost

---

[8]Given the unfortunate history of unethical medical experiments run on Black Americans, racialized attitudes towards the Covid-19 vaccine is not too surprising.

[9]Created by Freund and Schapire in 1997.

learns to correct its own mistakes, and makes relatively few assumptions about the data, making it a good "out-of-the-box" classifier.
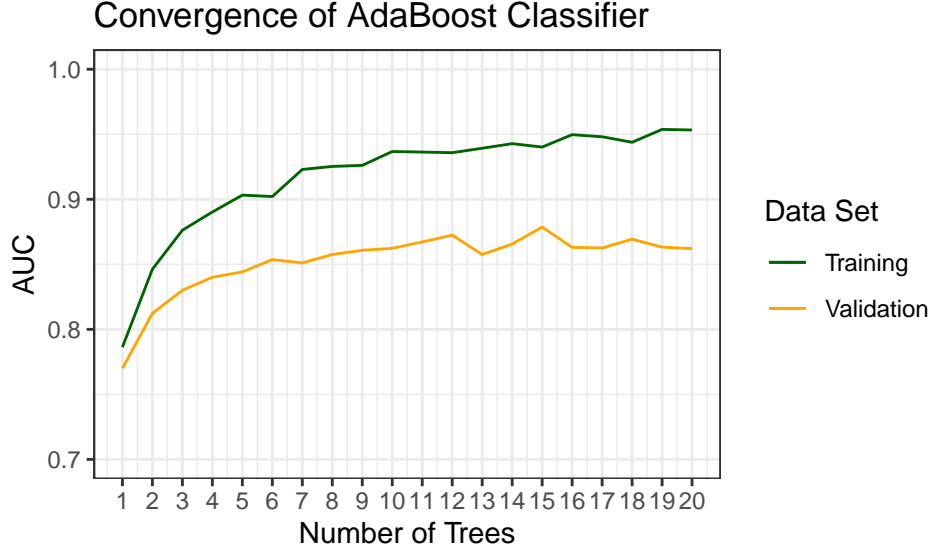

Figure 3: AdaBoost AUC Score Convergence

We trained our AdaBoost model on the following county-level demographic features: (1) percent college educated, (2) total population, (3) percent Black, (4) percent working in production, (5) income per capita, (6) poverty rate, (7) percent of walking commuters, and (8) Trump 2020 vote share. We denote the number of trees in Adaboost, the algorithm's main parameter, as m, and plot the AUC scores on training and validation sets as m ranges from 1 to 20. As we would expect, the model overfits the training set somewhat. As the number of trees increases past twelve, validation AUC seems to oscillate around a plateau, whereas training AUC keeps increasing. We decided to set $m = 6$, as validation AUC drops slightly at $m = 7$, and overfit starts getting significantly worse for larger m. This relatively simple model achieved an average accuracy of 82.9% on the test data set as well as an AUC score of 0.90.

The AdaBoost model depended mostly on the 2020 election results, percentage of Black population, and income per capita in that order, these features accounting for 76% of overall variable importance. The table below shows the importance percentage for each of the eight predictors.

Feature	% Importance
% Black	23.7
Income per Capita	16.5
% College Educated	4.9
% Trump Vote	36.0
Poverty Rate	6.9
% Prod. Workers	1.3
Total Population	6.5
% Walking to Work	4.2

3.2 Random Forest

In recent years, tree-based methods have exploded in popularity due to its flexibility, ability to learn non-linear relationships, and excellent prediction accuracy. As a non-parametric classification algorithm, it makes no formal distributional assumptions about the data. Furthermore, random forests have a natural way to rank

4

the importance of variables. The variable importance is determined by the mean difference in out-of-bag prediction error between the original data set and the data set with said variable permuted; the score is normalized by the standard deviation of the differences. We used the `randomForest` package in R to implement the algorithm.
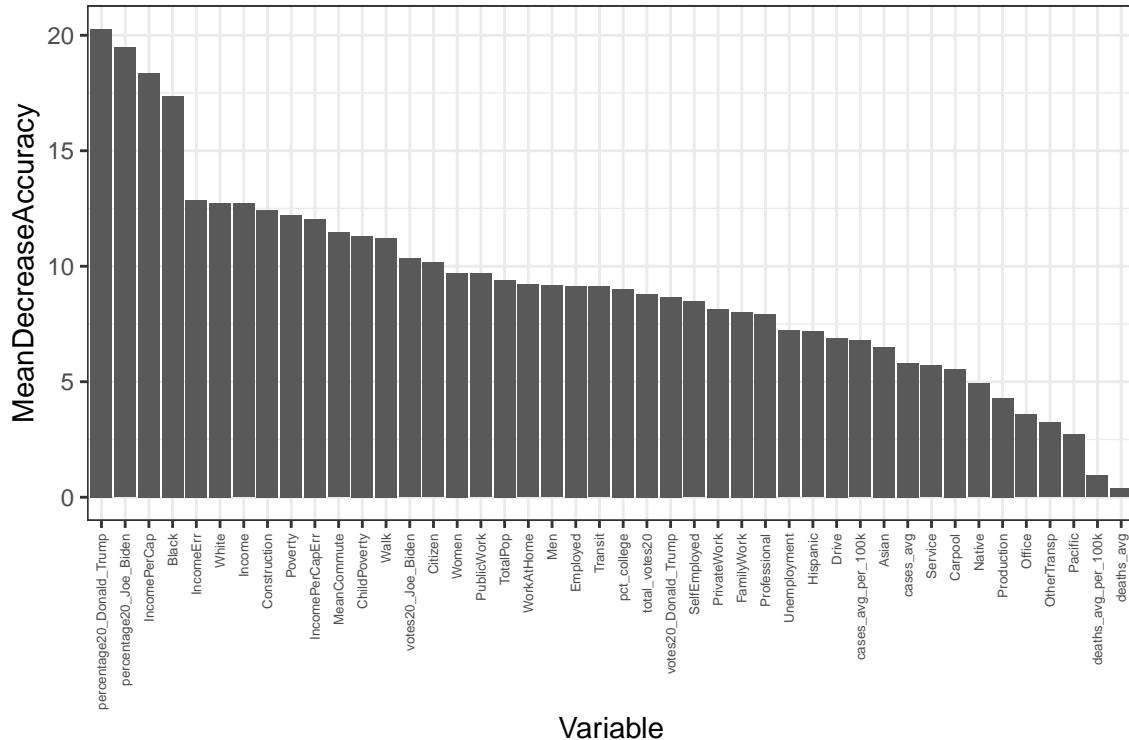

Figure 4: Feature importance for random forest classifier.

The random forest classifier identified four major factors using the mean decrease in prediction accuracy criteria. These four are the proportion of people who voted for Joe Biden and Donald Trump in 2020, income per capita, and proportion of people who are Black. Clearly, the source of vaccine hesitancy is multi-faceted. The four major factors show that political views and socioeconomic status play roles in mistrust of the vaccine.

3.3   Logistic Regression

We explored whether or not a parametric model could accurately explain our data using logistic regression. This model assumes that the log odds of the probability that a given county is vaccine-hesitant can be written as a linear combination of our predictor variables, i.e.

$$\log \frac{P(Y=1)}{1-P(Y=1)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p. \tag{1}$$

The parametric constraint may not accurately describe our data, so it will be interesting to examine how well this model performs compared to more "black-box" models like Random Forest and Adaboost, both of which make no implicit assumptions on the data generating process.

We have many degrees of freedom in the choice of which subset of predictors $(X_1, \ldots, X_k)$ to use in our model. Rather than choosing a subset arbitrarily or naively (e.g., the set of all predictors), we adopt a principled approach using the forward stepwise and backward stepwise model selection process, implemented

in R. The forward stepwise approach begins with the empty (intercept-only) model for the log odds, and at each steps adds in the predictor that results in the largest decrease in model AIC, terminating when no further improvement is achieved. The backward stepwise process does the same, but begins with the full model (every predictor variable) and removes a predictor at each step to increase the AIC. Additionally, we center and scale all predictors prior to the model fitting and selection process so that the coefficients β_i will be on comparable scale.

After running both procedures, we can examine the variables included in each model. Below, we plot the AIC of the forward stepwise model at each step, along with the predictor added at each step. We see that the final model chosen by this procedure includes 22 total predictors.

Forward Stepwise Procedure – Logistic Regression
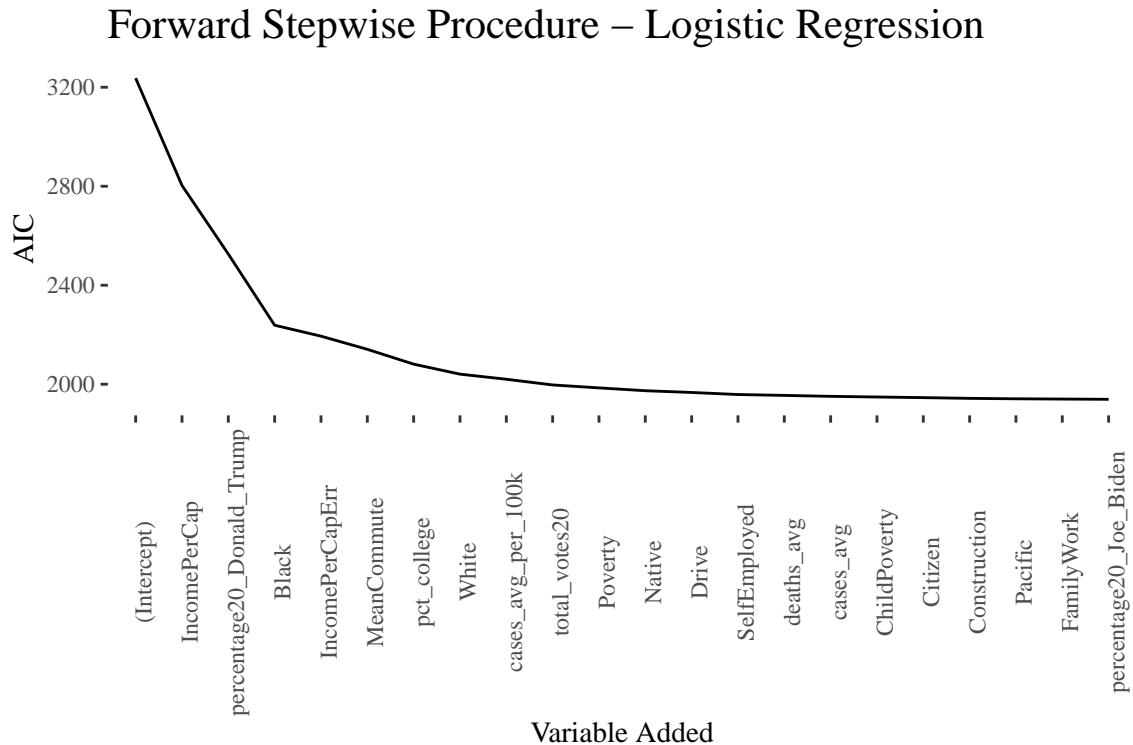
It's less informative to examine a similar plot for the backward stepwise procedure, as we would see the variables removed rather than those which remain. Instead, we examine the plots for the magnitude of the logistic regression coefficients for each of the two models. For the forward stepwise model, predictors are plotted in the order in which they were added to the model. For the backward stepwise model, the predictors are in arbitrary order.
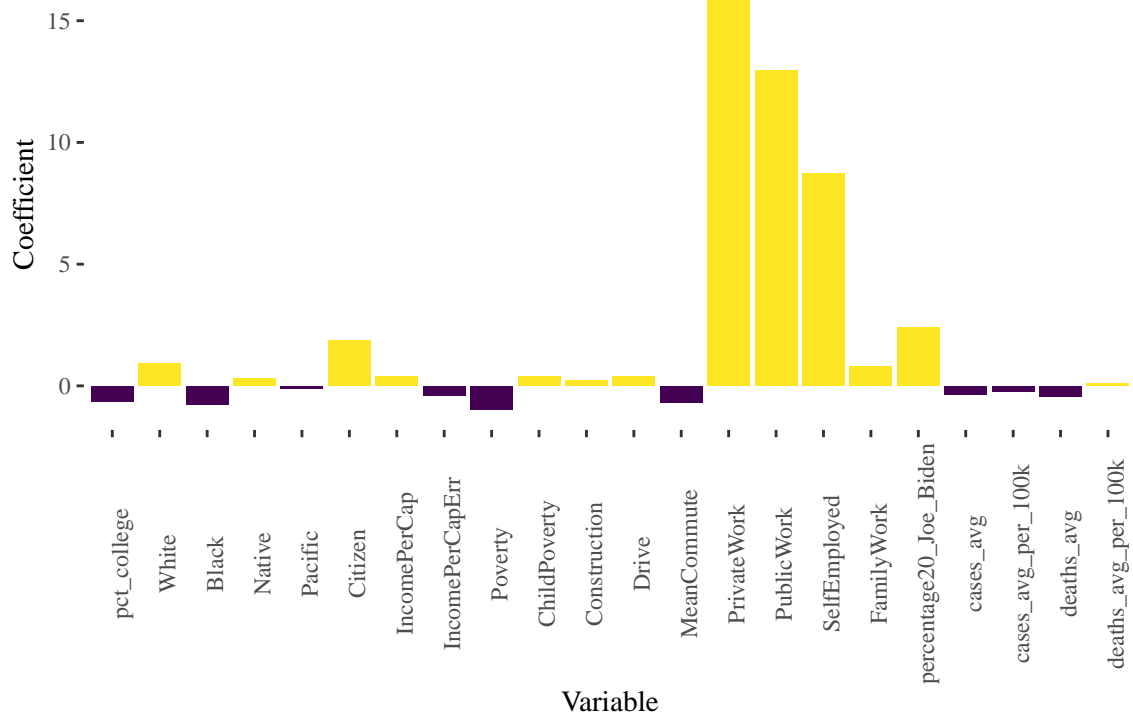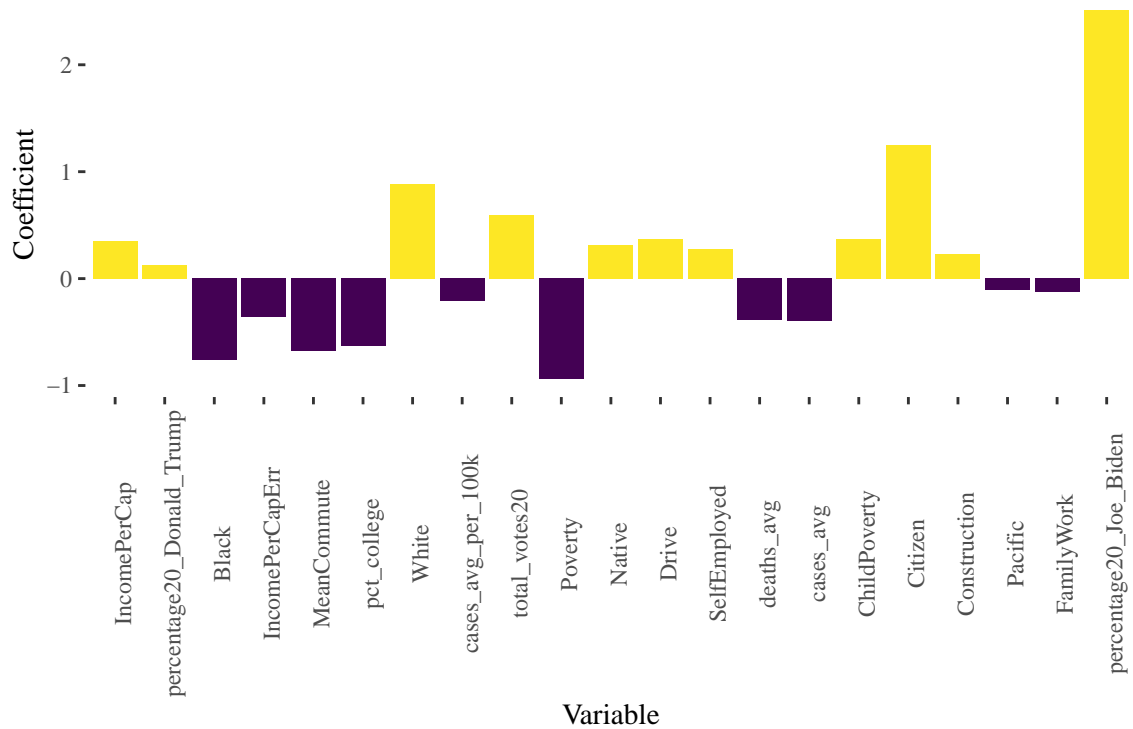
Because the data are standardized, the scale of the coefficients are comparable, allowing us to use these coefficients as surrogates for variable importance in the prediction problem. Larger coefficients in magnitude indicate a feature which a higher impact on the log odds of a county being vaccine hesitant - large positive coefficients indicate that a county with higher values for a particular predictor are *more* likely to be vaccine hesitant; large negative coefficients indicate that large values for that predictor are likely to indicate a county

that is not vaccine hesitant.

Keep this interpretation in mind, the results are somewhat surprising, particularly politically. Naively, one might expect counties with larger share of democratic voters (higher values for \texttt{percentage20_Joe_Biden}) to be less vaccine hesitant, yet rather surprisingly our data suggests the opposite. A large positive coefficient on this \texttt{percentage20_Joe_Biden} indicates that a county with a higher than usual value for this field is more likely to be vaccine hesitant. Of course, interpretation is not so simple: we also see a positive coefficient on \texttt{percentage20_Donald_Trump} which runs counter to the intuition that the coefficients on these two predictors should be different signs. Part of the difficultly in interpreting these numbers is that the coefficients should be considered as the impact of the predictor *after controlling for the effects of the others*. As some of the predictors are highly correlated with one another, this intricacy may make interpretation more difficult.

That being said, many of the results are still intuitive. We see in both models a negative coefficient on the percentage of people who are college educated, suggesting that education makes people in general less vaccine-hesitant. Counties which have in general been hit harder (either overall, or presently) with COVID-19 seem to be less vaccine-hesitant as well, as indicated by the negative coefficients on \texttt{cases_avg, cases_avg_per_100k, deaths_avg} in both models.

Characterizing the other predictors is somewhat difficult, but in general we speculate that many can regarded with respect to how urban a particular county is. Increased rates of office work, per capita income, poverty, and commute time all seem to suggest a county is less likely to be vaccine-hesitant, and intuitively we tend to associate these attributes with more urban living. Some attributes associated with more rural living (e.g., a lack of racial diversity in the form of a high percentage of "White" residents) on the other hand tend to suggest increased vaccine hesitancy.

3.4   Comparison of models / assessment of errors

```
## [1] 0.4140263
```

4   Some sort of post-hoc analysis

5   Conclusion

6   Bibliography