Predicting County-Level COVID-19 Vaccine Hesitancy

Prayag Chatha, Josh Jiang, Declan McNamara

December 09, 2021

1 Introduction

As of December 2021, about 59% of the US population has been fully vaccinated against COVID-19. Despite America's wealth, advanced medical sector, and its creation of the first vaccine, America ranks just 19th in the world in terms of vaccination rates.[1] Given the abundance of domestic vaccine doses, it is reasonable to assume that large segments of the American populace are unwilling to take the vaccine. Media narratives attribute this public hesitancy to many causes, ranging from ranging from political polarization to consumer anxiety about hidden costs.

In this study, we examine the relationship between county-level vaccination rates and local demographics, such as racial composition, economic indicators, and political tendency. A county-level study can reveal interesting geographic patterns by aggregating the behavior of individuals who tend to belong to the same social unit (e.g. socioeconomic class or race). A quantitative analysis of what factors predispose communities toward or against vaccination may be useful in crafting effective public health messaging. # Data

1.1 Data Sources

We derived our data on county-level vaccination rates[2] from a scrape of the CDC's county view dashboard dating back to June 20th, 2021. We chose this date as roughly half of the US population had been fully vaccinated by this point in time.[3]. For potential predictors of county-level vaccination, we obtained county-level results for the 2020 US Presidential election[4], demographic data (e.g. racial composition, income levels) from the 2015 5-year Census estimate[5], as well as county education levels (i.e. prevalence of a bachelor's degree or higher).[6].

1.2 Data Overview

Our combined data set contains complete data on 2,809 counties[7] in 47 states (sans Alaska, Hawaii, and Texas) as well as Washington, DC. These counties ranged from metropolitan areas[8] such as Los Angeles, Cook, and Maricopa counties to rural areas containing only a few hundred inhabitants. As counties primarily represent geographic area, and people tend to be clustered in urban areas, most counties are less populated 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://ourworldindata.org/covid-vaccinations
[2] https://github.com/covidestim/cdc-vaccine-data
[3] https://ourworldindata.org/covid-vaccinations?country=USA
[4] https://www.kaggle.com/etsc9287/2020-general-election-polls
[5] https://www.kaggle.com/muonneutrino/us-census-demographic-data
[6] https://www.ers.usda.gov/data-products/county-level-data-sets/
[7] Out of the approximately 3,100 counties and equivalent municipalities in the USA
[8] In New York City, vaccination data for all five boroughs were combined, which led to its exclusion from the final data set.

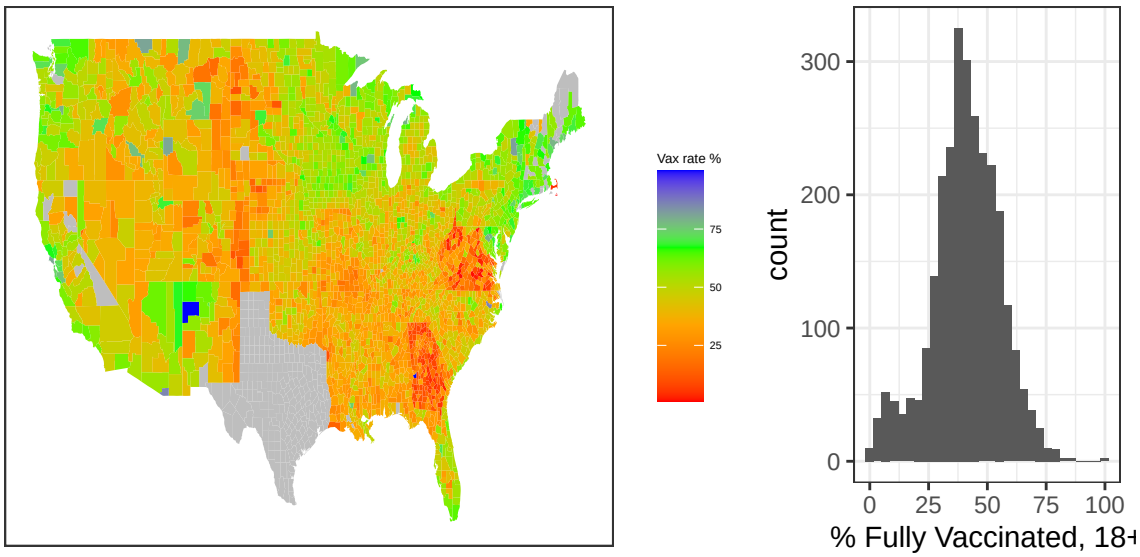


Figure 1: Map and Distribution of 18-Plus vaccination Rates by County, June 20th, 2021

1.2.1 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. Also near the top of the list were several affluent counties, such as Montgomery MD, Marin CA, and Glacier MT, which all had a vaccination rate of 80% or more. The former group could indicate an effect where counties that were particularly hard-hit by Covid-19 (i.e., poor and less-white areas) have seen widespread vaccination adoption. The latter group suggests that wealth correlates with vaccination willingness. In general, the counties with the highest vaccination rates generally leaned Democratic in the 2020 presidential election.

1.2.2 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. In today's politicized climate, Republicans may be more likely to be skeptical of a vaccine rollout happening under a Democratic administration.

1.2.3 Geographic Patterns

Figure 1 shows a map of vaccination rates in the lower 48 states (minus Texas). THe southern states West Virginia, Virginia, and Georgia are home to many counties with some of the lowest vaccination rates. The Great Plains (e.g Nebraska, North Dakota) appear to have lower-than-average vaccination adoption, though rural counties in neighboring states Minnesota and Iowa tend to be widely-inoculated. The Northeast, the West Coast, and southern Florida (notably home to many seniors) are some other areas with widespread vaccination.

1.3 Exploratory Data Analysis

Another view of our data set involves comparing all counties' vaccination rates based on single predictors. Figure 2 shows scatterplots of vaccination against six features roughly corresponding to education, wealth,

political tendency, ethnicity, economic class, and population density: we speculated that these features have relatively high variance across American counties.

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). Despite an abundance of free vaccine doses, poorer Americans may be less able to miss work for a shot or else fear hidden costs (especially if they lack health insurance). 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 ($\rho = -0.25.$).[9] 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. 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 America) tended to vote for Joe Biden in the last election, so this effect could be merely a reflection of existing political polarization.

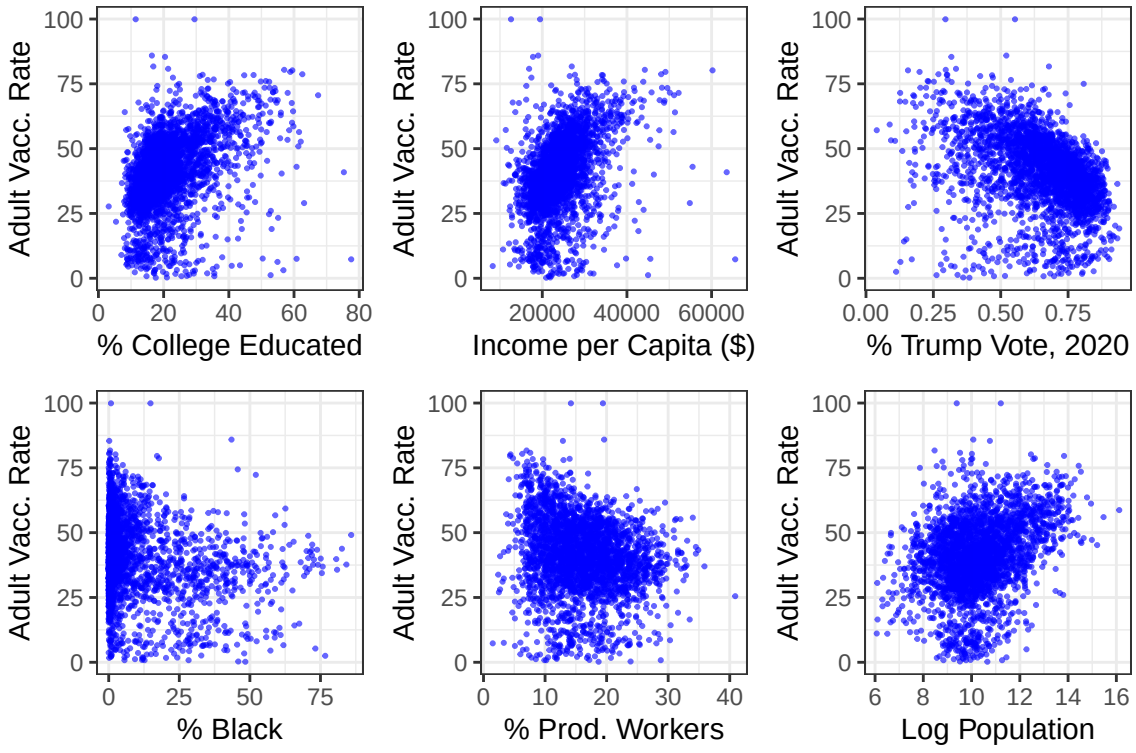


Figure 2: Correlation Scatterplots for Select Demographic Features

2 Classifying Low and High Vaccination Counties

The outcome variable we aim to model is the proportion of fully-vaccinated adults. As binary classification is a more tractable problem than regression (especially on the 0-1 range), we framed our task as identifying vaccine-hesistant and vaccine-willing counties. Hence, we split counties into two categories: 1,646 low

[9]Given the history of unethical medical experiments run on Black Americans, racialized attitudes towards the Covid-19 vaccine is not surprising. In contrast, the percentage of white population was largely uncorrelated with vaccination.

vaccination counties having a rate less than 44%, and 1,163 high vaccination counties with a rate greater than or equal to 44%. This resulted in a 59-41 split in the data. It is worth noting that about 70 million Americans live in "low" counties and 205 million Americans live in "high" counties.

We compare the performance of three classification models: AdaBoost, Random Forests, and Logistic Regression. We use a 75-10-15 split for training, validation, and testing sets. The purpose of the validation set is to select parameters that generalize well to unseen data with minimal overfitting, while the final performance is assessed on a previously-unseen test data set.

2.1 AdaBoost

AdaBoost, short for "adaptive boosting," is the classical[10] boosting algorithm for classification. The philosophy of boosting is that an ensemble of weak learners can act in concert as a strong learner. At each iteration of training, AdaBoost assigns additional weight to previously-misclassified observations, which helps it learn to correct its own mistakes. AdaBoost makes relatively few assumptions about the data, making it a good "out-of-the-box" classifier.

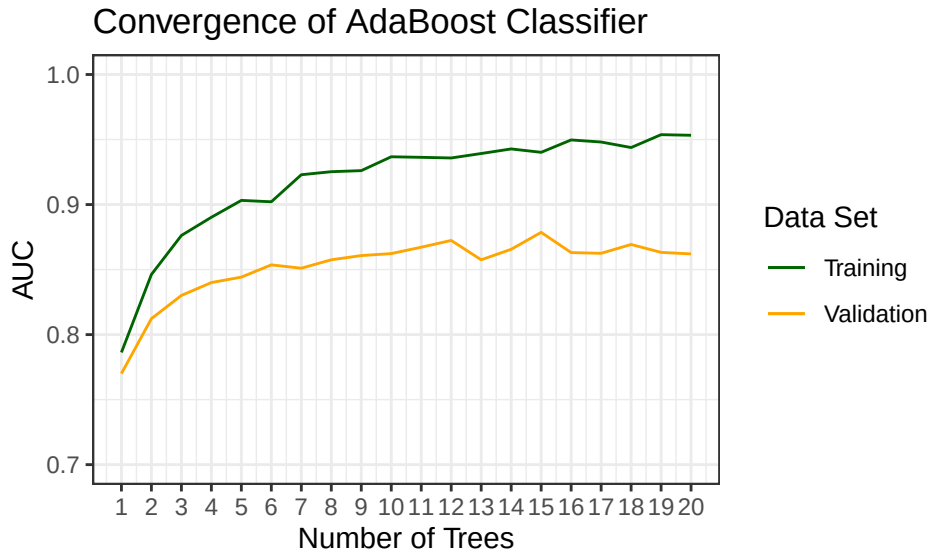


Figure 3: AdaBoost AUC Score Convergence

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.

[10] Created by Freund and Schapire in 1997.

Table 1: AdaBoost Feature Importance

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

2.2 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 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.

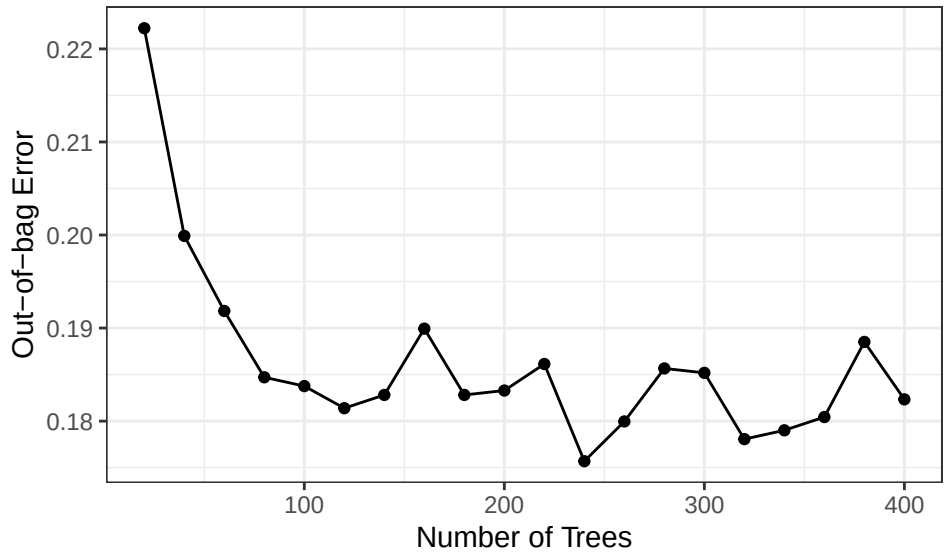


Figure 4: Out-of-bag Error Rate for Random Forest Classifier.

Similarly to AdaBoost, the random forest algorithm trains an ensemble of weak classifiers to become a strong classifier. We chose the number of trees in the forest, B, to be the one that minimizes out-of-bag (OOB) error. We decided to use $B = 240$, although we observed that OOB error seems to bottom out after $B = 100$. The algorithm was able to achieve a 85.5% prediction accuracy on the test set.

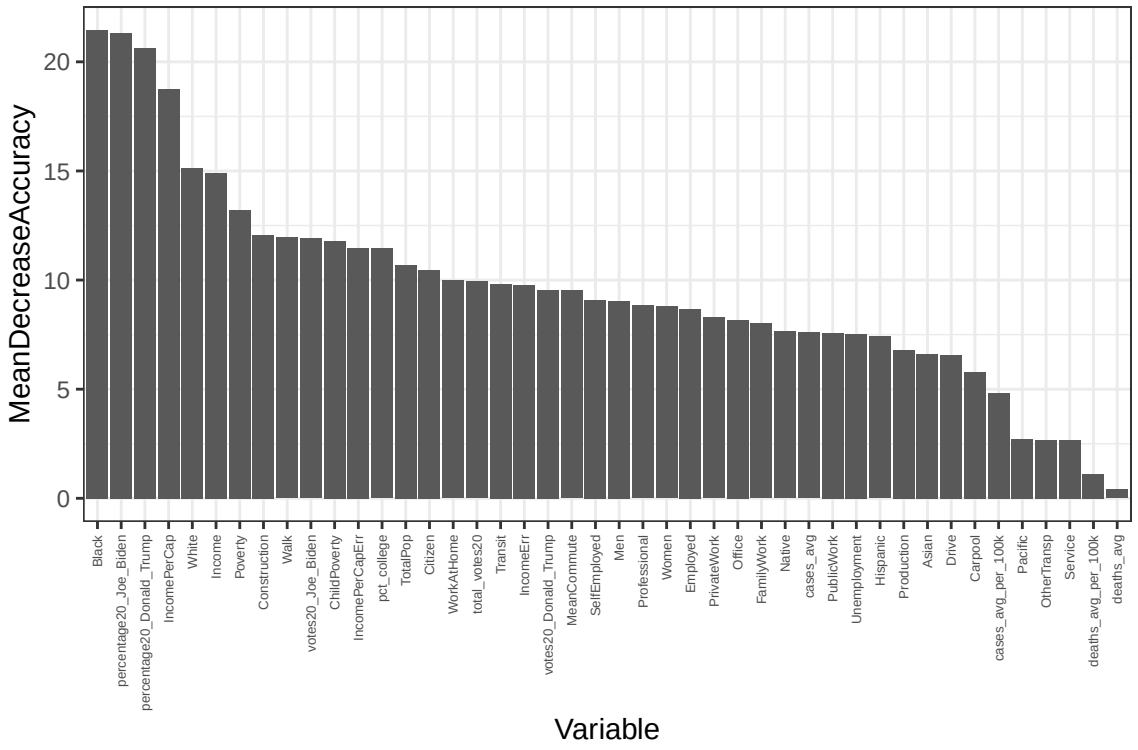


Figure 5: Feature importance for random forest classifier.

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 multifaceted. The four major factors show that political views and socioeconomic status play roles in mistrust of the vaccine.

2.3 Logistic Regression

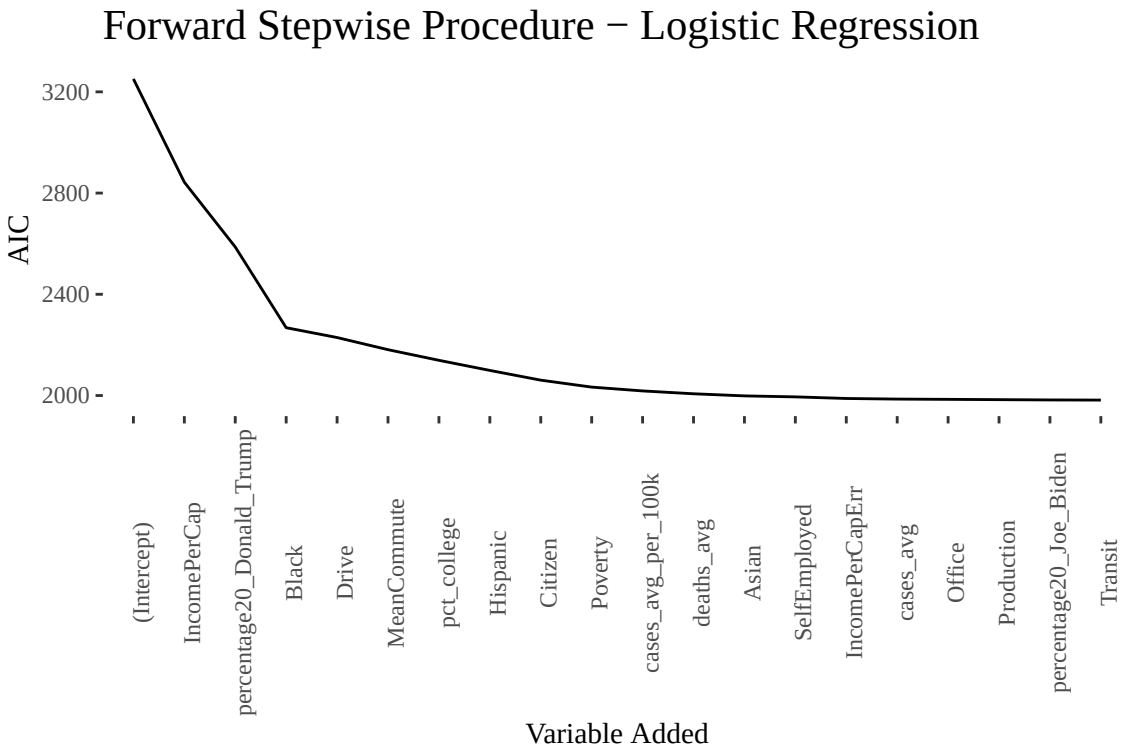
We explored whether or not a parametric model could accurately explain our data using logistic regression. This model assumes that the log odds of the probability that a given county is vaccine-hesitant can be written as a linear combination of our predictor variables, i.e.

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

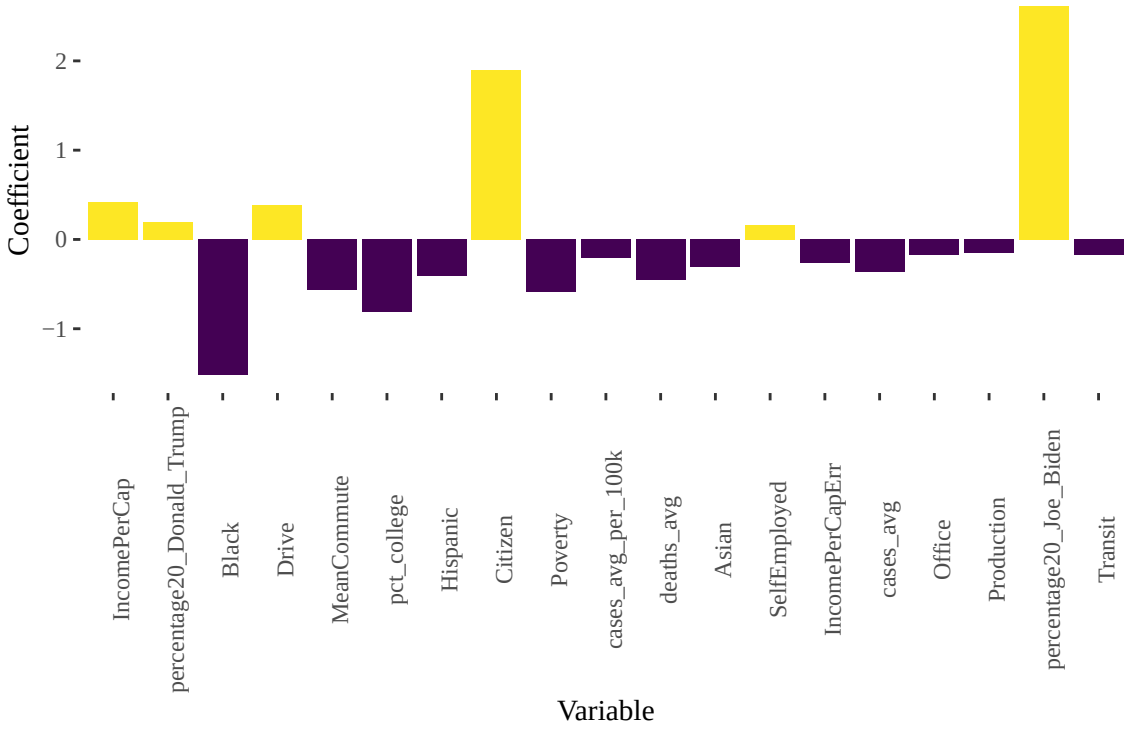
The parametric constraint may not accurately describe our data, so it will be interesting to examine how well this model performs compared to more "black-box" models like Random Forest and Adaboost, both of which make no implicit assumptions on the data generating process.

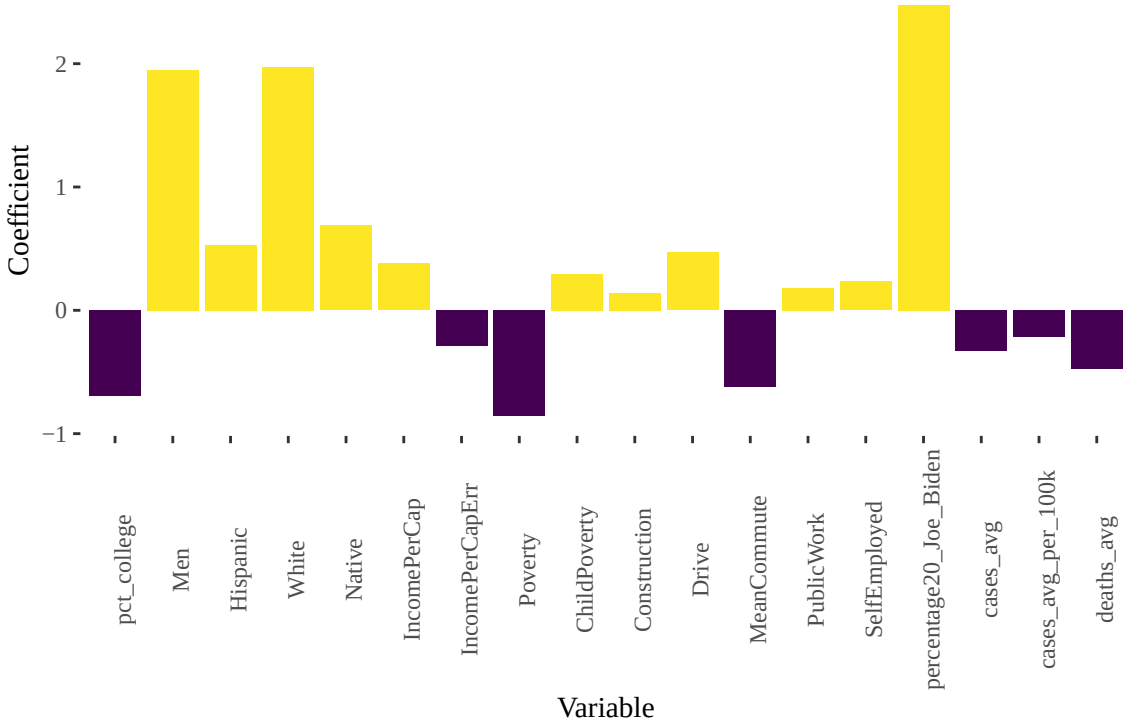
We have many degrees of freedom in the choice of which subset of predictors $(X_1, \ldots, X_k)$ to use in our model. Rather than choosing a subset arbitrarily or naively (e.g., the set of all predictors), we adopt a principled approach using the forward stepwise and backward stepwise model selection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



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 less clear.

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 we speculate that many are proxy measures of how urban a particular county is. Increased rates of office work, per capita income, 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 city life. 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.

Having fit these two models on our training data, we can examine their test accuracy. On the test set, the logistic regression model fit with the forward stepwise procedure had a test accuracy of 84.6%, while the model fit with the backward stepwise procedure had a test accuracy of 85.3%. The similarity of the two accuracies suggests that the two models are substantially similar, even though they do differ in a few predictors.

2.4 Model Comparison

The three models were trained on the same training set. To assess their performances, we used the same test set for each model.

Table 2: Mean Accuracy on Test Data

Model	Test Accuracy
AdaBoost	.829
Random Forest	.855
Logistic Regression	.853

Table 2 shows the prediction accuracies on the test set for the three classification models. Since logistic regression produces a linear decision boundary while random forest can learn more complex ones, we found it surprising that logistic regression performed on par with random forest. The relatively poor performance of AdaBoost is likely due to it only training on a subset of the predictors.

3 Post-Hoc Analysis

Since random forest is the best classification model considered, we will perform some post-hoc analysis on its performance. In particular, we are interested in identifying geographic patterns in the model's classification error. To allow ourselves as complete a map of the US as possible, we decided to use the entire data set in this analysis. For a given observation x_i in the training set, we predicted it using only trees not trained on the observation. Of course, while the results of the training set will not reflect the true capabilities of the random forest, we believe it will be close enough.

Table 3: Confusion matrix for random forest. FP=0.190, FN=0.173.

	Vaccine Willing	Vaccine Hesitant
Positive	861	202
Negative	302	1444

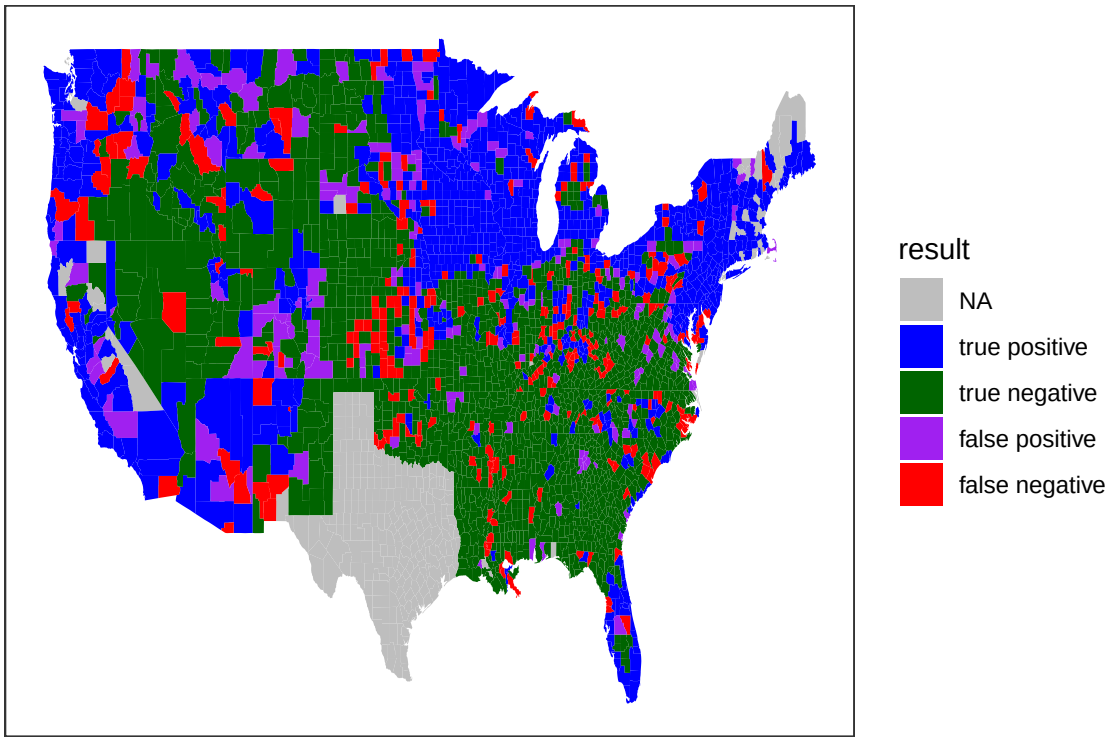


Figure 6: Map of counties color coded by classification result using random forest.

There appears to be a concentration of misclassifications at the border of hesitancy regions. For example, there are many false negatives in parts of Illinois, Indiana, Ohio, Kentucky. Those states are known to be the boundary between the Midwest/Northeast, which are generally vaccine willing regions, and the South, which contains many of the most vaccine hesitant areas.

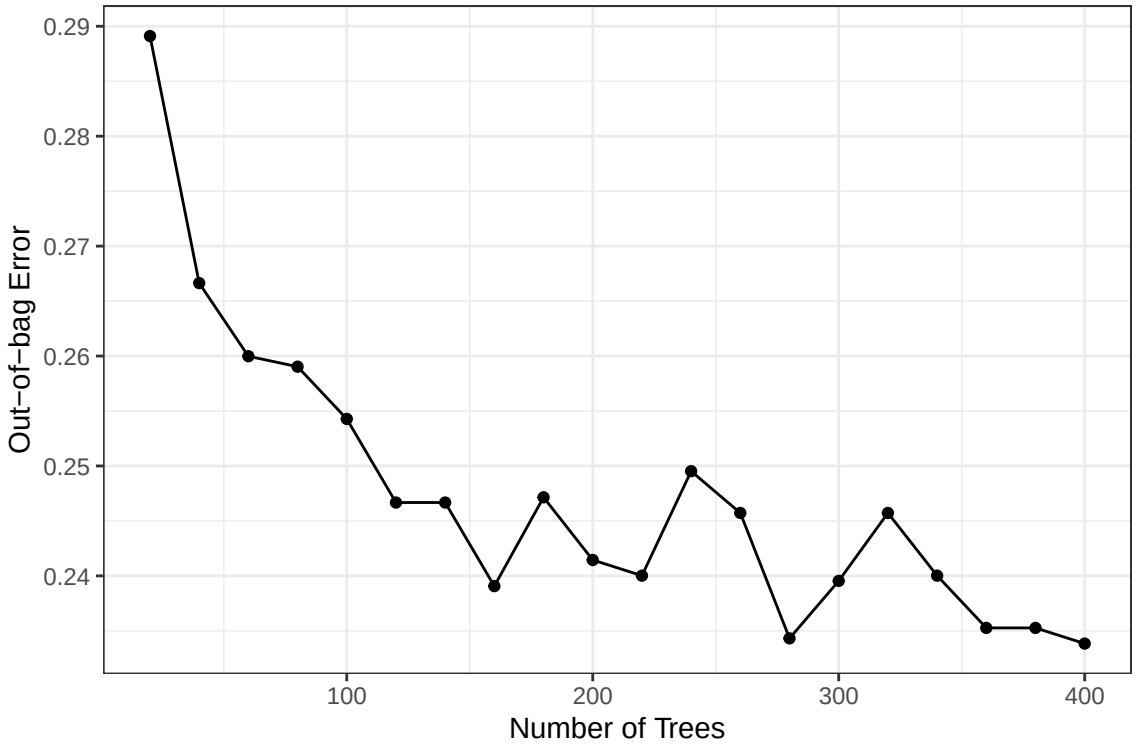
In table 4, we looked at the misclassifications on a county level. In particular, we looked at the counties that had the worst misclassifications - highly vaccinated counties that were classified vaccine hesitant and vice versa.

Table 4: Worst misclassifications - false positives and false negatives - and their characteristics.

	County	Vax Rate	Joe Biden	Income per Capita	Proportion Black
Worst FP	Nantucket, MA	0.012	0.717	45000	0.093
2nd Worst FP	Roanoke, VA	0.019	0.388	31370	0.056
3rd Worst FP	Winchester City, VA	0.036	0.549	26182	0.104
Worst FN	Martin, NC	0.859	0.471	19032	43.5
2nd Worst FN	Santa Cruz, AZ	0.854	0.672	17795	0.2
3rd Worst FN	Graham, KS	0.750	0.171	28233	0.031

4 Stability Check

Having concluded our analysis, we would like to verify that our results are robust to the time frame of data collection. Recall that we only selected a particularly snapshot in time, right around June 20th, 2021, to determine vaccine hesitancy. In this section, we consider a different snapshot in time, the week of April 18th, 2021, and rerun our analysis for our best model, random forest. We will examine whether or not our ability to predict vaccine hesitancy remains stable, and whether or not the variables most relevant to predicting vaccine hesitancy will remain similar as well. Due to increased vaccination rates over time, this analysis calls for a reformulation of how we label the ground truth of "hesitant" or "not hesitant" for each county. We again choose the 60th quantile of vaccination rate as the cutoff value; those counties above this threshold (29.7%) will be labeled as vaccine-accepting, while those below will be labeled vaccine hesitant.



We see some differences compared to the OOB error rate when using the snapshot from June, primarily in the error rates. While in June these out-of-bag error rates got down to as low as 18% or so, when we train on the data set from April our error is a bit higher, around 23% at best.

In trying to determine what factors primarily contribute to vaccine hesitancy, consistency across these two timeframes will help convince us that some variables truly are reliable predictors, rather than black box aids for the model. Below, we plot the feature importances across both time frames, June and April, for comparison.

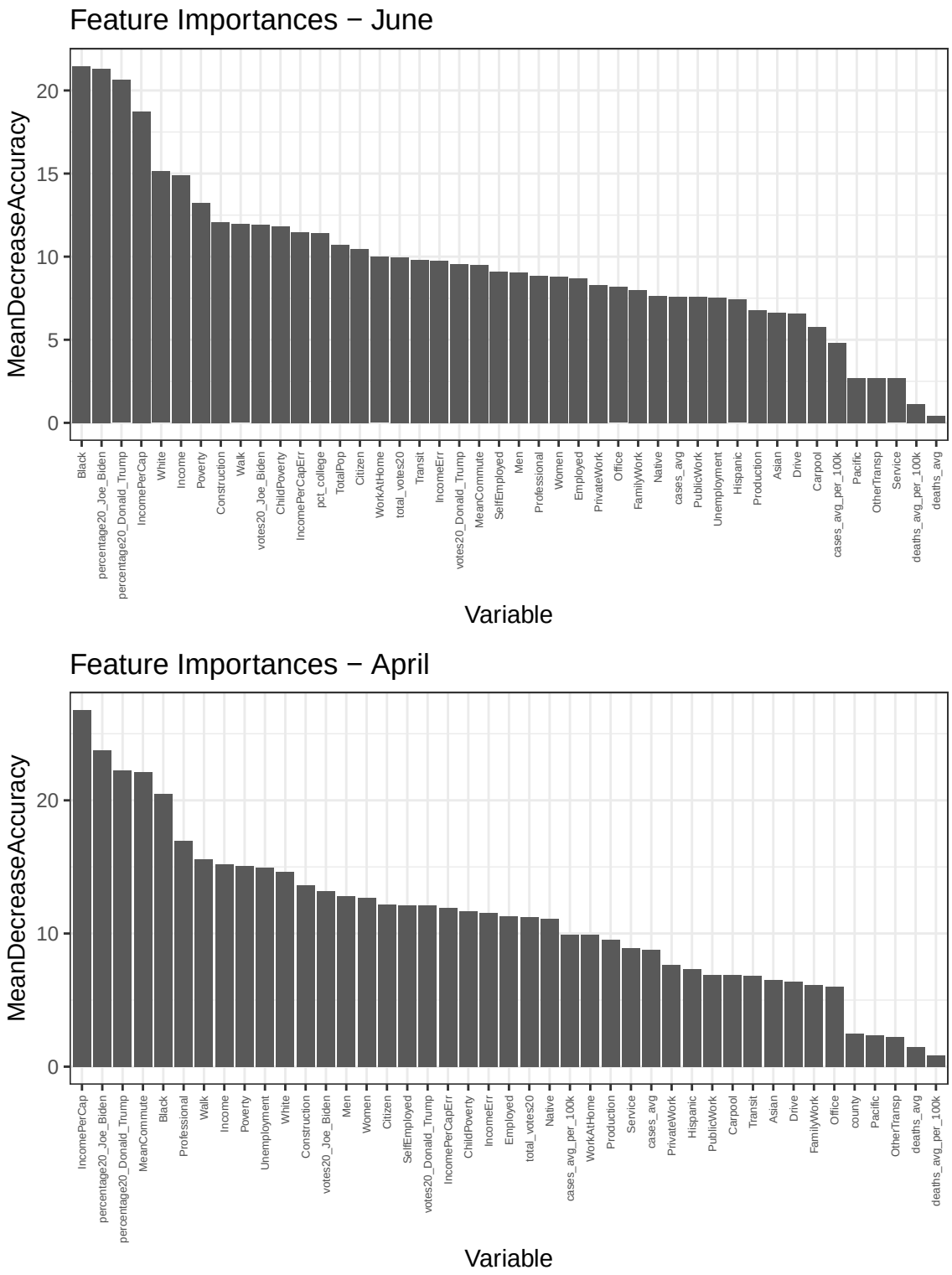
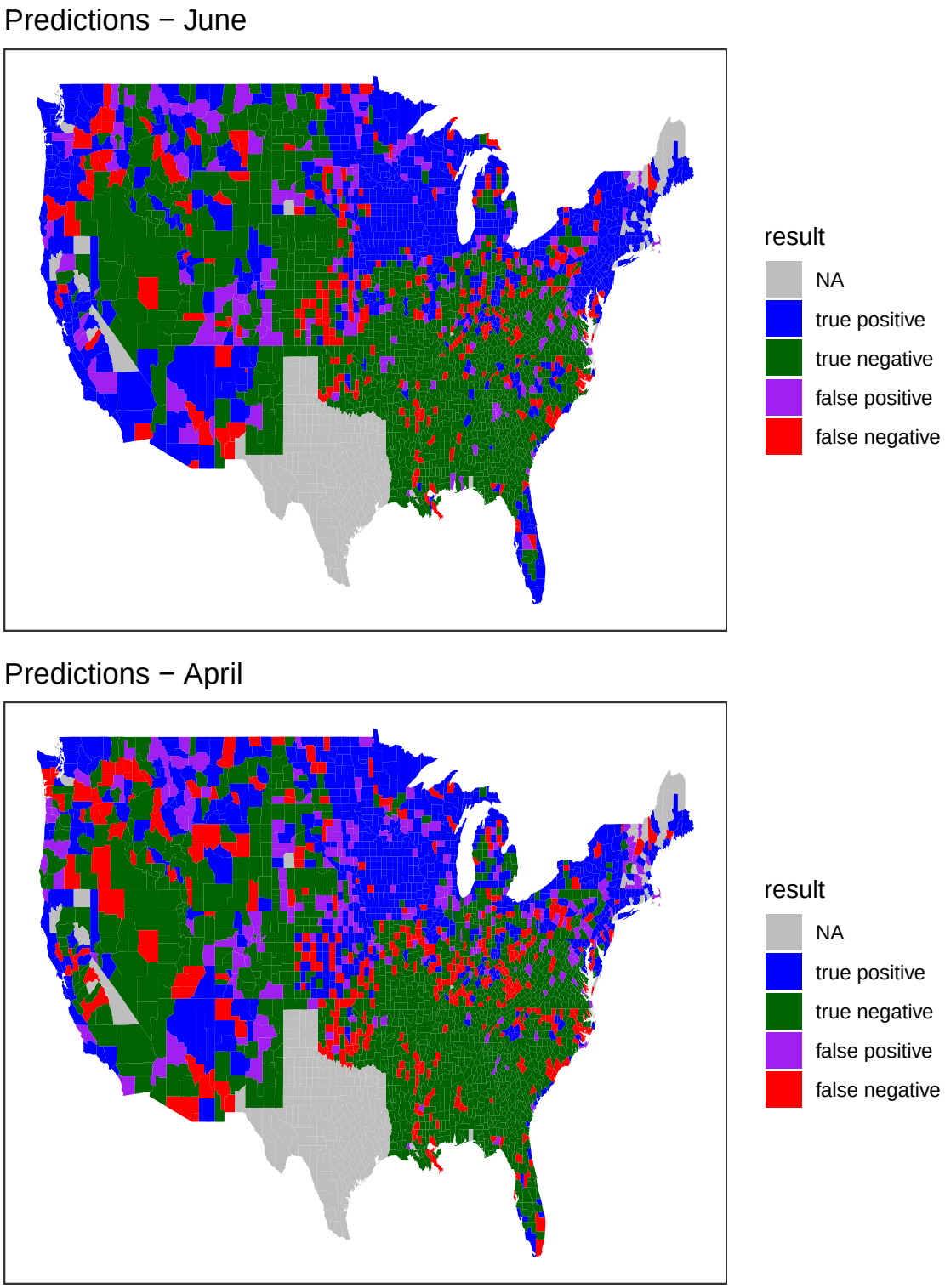


Figure 7: Feature importance for random forest classifier.

While we see some differences, there are striking similarities as well. Both plots indicate a steep drop-off in importance after the first five predictors, suggesting that these five are most important to model beyond

others. In both time frames, we have consistency in four out of the top five predictors: income per capita; percentage of votes for Donald Trump in 2020; percentage of votes for Joe Biden in 2020; and percentage of black residents. While there order are shuffled slightly within the top five importances, seeing consistency in these feature importances is encouraging - it reassures that us our findings are valid rather than happenstance.

We may also be interested in exploring the question of how our predictions change over time for each county. Below, we plot our model predictions from both April and June.



Overall, the model errors seem consistent. The same counties often seem to be missclassified in both time frames, suggesting that these few may simply be not well explained by the model or predictors in some fashion. Interestingly, though, we see that our model is indeed capable of catching changing behaviors over time. Florida in particular demonstrates this - we see a high level of vaccine hesitancy in April in the form of lots of green counties (vaccine-hesitant counties), which we tend to predict correctly. By June, though,

many counties in upper Florida have switched from vaccine-hesitant to not so, indicated by the presence of blue. Interesting, our model is able to follow the change in trends in these counties, despite seemingly relying on some fields which are static over time: percentage of votes for a particular candidate; income levels; demographics. We speculate that our inclusion of COVID-19 related data allowed us to follow changing behavior because this data causes people to change the vaccine views - as counties are hit hard with the pandemic, people in those counties may be more inclined to get the vaccine where they might not have before.

5 Conclusion

6 Bibliography