Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Machine Learning with Applications 10 (2022) 100408

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Predictors of COVID-19 vaccination rate in USA: A machine learning approach

Syed Muhammad Ishraque Osman [a,1], Ahmed Sabit [b,*]

[a] Jack Welch College of Business & Technology, Sacred Heart University, West Campus, East Building - 1st Floor, 3135 Easton Turnpike, Fairfield, CT 06825, United States of America

[b] Department of Biostatistics, The Johns Hopkins University, 615 North Wolfe Street, Baltimore, MD 21244, United States of America

ARTICLE INFO

Keywords:
COVID-19
Vaccination
Vaccine hesitancy
Machine learning
Decision tree
Health policy

ABSTRACT

In this study, we examine state-level features and policies that are most important in achieving a threshold level vaccination rate to curve the effects of the COVID-19 pandemic. We employ CHAID, a decision tree algorithm, on three different model specifications to answer this question based on a dataset that includes all the states in the United States. Workplace travel emerges as the most important predictor; however, the governors' political affiliation (PA) replaces it in a more conservative feature set that includes economic features and the growth rate of COVID-19 cases. We also employ several alternative algorithms as a robustness check. Results from these checks confirm our original findings regarding workplace travels and political affiliation. The accuracy under different model specifications ranges from 80%–88%, whereas the sensitivity is between 92.5%–100%. Our findings provide actionable policy insights to increase vaccination rates and combat the COVID-19 pandemic.

1. Introduction

The COVID-19 (SARS-CoV-2) virus, first detected in December 2019, quickly spread all over the world and took the lives of almost six million people with approximately 436 million confirmed cases (Dong, Du, & Gardner, 2020). The Centers for Disease Control and Prevention (CDC) reported the first COVID-19 case in the United States on January 20, 2020.[2] Since then, over 78 million people got infected and over [945,000] people died.[3] COVID-19 disrupted our everyday life and devastated economies across the world. As a result, returning to "normal" primarily through mass vaccination has become a priority in the U.S and across the world. Several vaccines are now in use, developed in an unprecedented effort to combat the pandemic. As the U.S is pushing towards returning to "normal", getting more people vaccinated has become a policy priority. In the second week of December 2020, the Food and Drug Administration (FDA) granted an emergency use authorization (EUA) for the Pfizer-BioNTech vaccine. The Moderna vaccine and the Johnson and Johnson vaccine got the same EUA on December 18, 2020, and February 22, 2021, respectively.[4] After that, the authorities started to rollout vaccines across the U.S and took several aggressive policy measures to encourage mass vaccination. To date, [64.9%][5] of the total population in the U.S are fully vaccinated. While vaccination rates in some states are impressive, there are twenty-one states where fewer than 60% of the population has been completely immunized. States with low vaccination rates often experience higher COVID-19 cases (Borchering et al., 2021), also several reports show that the unvaccinated population is more vulnerable against the other variants such as the highly contagious delta variant (Dyer, 2021). In order to address the low vaccination rates, 26 states announced different financial and non-financial incentives. However, studies, such as (Sabit, Ahmad, & Abdul Baten, 2022; Walkey, Law, & Bosch, 2021), find that incentives are not effective enough to encourage people to get vaccinated. For example, Ohio stopped its lottery program, stating that the program's effect was "short-lived" (Huggins, 2021). Against this backdrop, there is a renewed interest to know what factors most effectively predict the

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail addresses: osman.datascience@gmail.com (S.M.I. Osman), sabit006@gmail.com (A. Sabit).

[1] First author.

[2] https://www.cdc.gov/media/releases/2020/p0121-novel-coronavirus-travel-case.html

[3] https://covid.cdc.gov/covid-data-tracker/#cases_totalcases

[4] https://www.cdc.gov/museum/timeline/covid19.html

[5] https://covid.cdc.gov/covid-data-tracker/#vaccinations_vacc-total-admin-rate-total

https://doi.org/10.1016/j.mlwa.2022.100408
Received 3 November 2021; Received in revised form 2 September 2022; Accepted 2 September 2022
Available online 16 September 2022

vaccination rates across the states in the U.S. In other words, what are the unique features of the high vaccination states that make them different than the states with low vaccination rates. We attempt to answer this question in this study.

One of the many ways to combat the COVID-19 pandemic is to achieve herd immunity through mass vaccination. According to several studies, (Randolph & Barreiro, 2020, Bartsch et al., 2020, Goldblatt et al., 2022) at least 60 percent to 70 percent of the total population should be vaccinated to obtain herd immunity. But most of the studies (Malik, McFadden, Elharake, & Omer, 2020, Viswanath et al., 2021) estimate a lower vaccination rate in the U.S than this estimated threshold. Data also shows that vaccination rate in 21 states are still under 60% threshold, whereas the rates in 38 states are still under 70%.[6]

Vaccine hesitancy and its role in lower vaccine uptake is not new and well documented in the literature. However, given that we cannot reduce vaccine hesitancy overnight, it is crucial to understand what other factors are instrumental in achieving a higher vaccination rate. In this context, examining the state-level policies and other state-level features that can play a critical role in vaccination uptakes is highly relevant.

1.1. Problem definition and our approach

Our goal is to identify the most critical features that can predict which states will meet the vaccination threshold. Therefore, we choose CHAID, a decision tree algorithm, as our primary modeling technique. CHAID is easier to understand, faster to train, and interpretation is much more straightforward; in contrast, Random Forest (RF) and XGBOOST, which combines multiple DTs, are challenging to interpret and less prone to overfitting (Prajwala, 2015). Furthermore, RF and Gradient Boosting, XGBOOST in our case, differ in how the DTs are created. Instead of creating DTs independently, XGBOOST algorithm creates them additively to improve on the deficiencies of the previous trees (Sagi & Rokach, 2021). In the robustness check section, we showed that the results from our CHAID-based models are robust to the results from both RF and XGBOOST algorithms; hence we choose CHAID, which is much simpler. The following three algorithms, ElasticNet (ECV), LASSO (LCV), and RIDGE regression (RCV) can also tackle the overfitting problem (Ranstam & Cook, 2018); however, they are not very useful in modeling non-linear relationships. Due to the branching structure of CHAID, a DT algorithm, it can model non-linear relationships relatively easily (Klosterman, 2019). CHAID also has superiority over Logistic Regression (LR) in terms of interpretation. Additionally, the non-parametric approach of CHAID does not force us to decide on the pre-assumed parameters like LR. To some extent, collinearity, missing values, and outliers in the data are easier to handle in CHAID (Tomaschek, Hendrix, & Baayen, 2018), whereas it can potentially affect LR coefficients. As with any regression model, overfitting is an issue with LR (Dreiseitl & Ohno-Machado, 2002); however, if the classes are not well separated, then DT algorithm can also potentially cause overfitting in the training data. This is why, apart from other algorithms, we also compare results from LR with our original CHAID-based models in the robustness section. Lastly, like CHAID, the Support Vector Machine (SVM) algorithm can also model non-linear relationships and handles outliers better (Chauhan, Dahiya, & Sharma, 2019). However, SVM tackles non-linearity using the "kernel trick" compared to hyperplanes in CHAID. As a result, we wanted to check if our results from CHAID-based models are comparable to those using the SVM algorithm.

Our study includes all 50 states in the United States. We used publicly available data from the Centers for Disease Control and Prevention (CDC), Johns Hopkins University COVID-19 data repository, Google Community Mobility Report (CCMR), Kaiser Family Foundation (KFF), and the federal reserve bank of Philadelphia. Our primary modeling approach is a Decision Tree algorithm (CHAID) with three different model specifications. As a robustness check, we also applied Random Forest and several other non-tree-based ML algorithms such as Lasso (LCV), Ridge (RCV), ElasticNet (ECV), Logistic Regression (LR), Support Vector Machine (SVM), and XGBoost (XGB).

Our study has several contributions to the existing literature. *First*, this is the first study that focuses on state-level variables in the context of the COVID-19 pandemic to understand the differential in vaccination rates across the states in the United States. *Second*, using several ML algorithms, we identify and rank the state-level aggregate factors that most effectively predict the vaccination rates. It will work as a benchmark for future research related to state-level factors. *Lastly,* in the context of a pandemic as severe as COVID-19, this study offers some actionable insights for policymakers on how to increase vaccination rates to curb the pandemic's effects effectively.

2. Related literature

The literature on vaccine hesitancy and the determinants of vaccination rate are rich. There are two broad focuses of this literature; first, to understand the socioeconomic factors that determine vaccine uptake. Second, to understand vaccine hesitancy and how the individuals' behavioral and psychological aspects play a role in vaccination decisions. Ruiz and Bell (2021) investigate intention to vaccinate using multiple regression analysis and show that race, gender, age, socioeconomic status, marital status, political affiliation, and news sources significantly predict vaccination intention. Khubchandani et al. (2021) used a representative sample in the United States and employed Chi square tests and Logistic Regression to find the significant predictors. They found that number of children at home and the probability of getting infected with COVID-19 are significant predictors in addition to what Ruiz and Bell (2021) found.

Burch, Lee, Shackelford, Schmidt, and Bolin (2022) use univariate logistic regression and find a more parsimonious set of variables (education, age and gender) that have high predictive power in making vaccination decision. They also find, like (Khubchandani et al., 2021), that the perceived risk perception about COVID-19 is critical when individuals are making deciding about vaccination. Cheong et al. (2021), on the other hand, examine if sociodemographic factors play a role in vaccination uptake. Using the XGBoost algorithm, they find that ethnicity, location, education, and internet access have the most predictive power.

Dror et al. (2020) explored a behavioral aspect by using a multivariate logistic regression model in a study among 1941 Israeli population. They find that the medical professionals who are not taking care of SARS-CoV-2 positive patients expressed a higher vaccine hesitancy. In another study, Yan, Lai, Ng, and Lee (2022) surveyed 1003 individuals in Hong Kong and used a hierarchical regression model to show that vaccine uptake by known others and trust in authorities, among others, are important predicting factors in predicting vaccine uptake. Mewhirter, Sagir, and Sanders (2022) used the Gradient boosting (GB) algorithm and also found the lack of trust in the COVID-19 vaccine, risk perception about the COVID-19 virus itself, along with age are the main predictors of vaccination decision. Using a different set of population and a estimation techniques, Lincoln et al. (2022) confirms that the overall mistrust surrounding the COVID-19 vaccine and the virus itself plays a critical role in vaccine uptake. They surveyed 2510 individuals across five high-income counties and used multifactorial logistic regression in addition to the Random Forest algorithm.

Some studies forecast the percent of the population willing to get vaccinated. Malik et al. (2020) estimated that around 67% of the U.S population is willing to accept a COVID-19 vaccine if it is recommended for them. However, they also pointed out noticeable geographic and demographic disparities in vaccine acceptance among the participants. In another study, Viswanath et al. (2021) used multivariate Logistic Regression and estimated that 65–68 percent of the population

[6] https://covid.cdc.gov/covid-data-tracker/#vaccinations_vacc-total-admin-rate-total

Table 1
Definition of the features.

Features	Description
Economic Index	A single statistic that summarizes economic condition of a state
Grocery Travels	Change in visits to grocery stores relative to the baseline period
Covid Growth Rate	Growth rate of COVID cases
Mask Mandate State	State issued mask mandate
Mask Mandate School	Requirement of wearing a mask in school
Park Visits	Change in visits to parks relative to the baseline
Political Affiliation	Political affiliation of the Governor of a state
Residential Travels	Change in visits to places of residence relative to the baseline
Retail & Recreation	Change in visits to places like restaurants, shopping centers and libraries relative to the baseline
Retail Sales	State-level retail sales
State Emergency	Whether a state declared emergency or not
Transit	Change in visits to transit stations (subway, bus/train stations) relative to the baseline
Vaccine Mandate State	Any type of vaccine mandate by a state
Vaccine Mandate School	Vaccine mandate for school employees
Workplace Travels	Change in visits to workplaces relative to the baseline

would accept the vaccine, which is significantly associated with risk perception about COVID-19. They also found that, like other studies mentioned above, sources of information, confidence in scientists, and political affiliation significantly affect the decision to get vaccinated.

However, most vaccination and vaccine hesitancy studies primarily focus on individual-level factors and used survey data. While individual-level factors are important, understanding the state-level variables are also important for effective policy making. Our study attempts to do this. Our study also did not use survey data. Issues with survey data, such as self-selection and over/understatement, are well documented (Starr, 2012) in the literature. As a result, our results are free from any such issue.

3. Data and variables

3.1. Data acquisition

We use data till July 01, 2021. We wanted to base our analysis on the early performances of the states in order to identify what state-level policies/features have the most predictive power in achieving the vaccination threshold. In the later months, many states, federal, and even county-level policies were formulated to accelerate the vaccination rates. Therefore, we believe that data from the early few months of vaccination would reveal more natural predictors that drive vaccination.

We used publicly available data provided by CDC (CDC, 2022), Johns Hopkins University (Dong et al., 2020),[7] Google (Google LLC, 2021) and Kaiser Family Foundation (KFF).[8] In particular, we extract vaccination data from CDC, COVID-19 cases data from Johns Hopkins, COVID-19 related community mobility data from Google,[9] and state level policy data from Kaiser Family Foundation website. We use COVID-19 Community Mobility Report (CCMR), which reveals how visits to areas like grocery stores, workplaces, and parks are changing across the country. This dataset illustrates how visits and duration of stay at various locations change over time compared to a baseline.

3.2. Target/outcome variable

Our main variable of interest is the state level of vaccination rate. We calculate vaccination rates as the percentage of the vaccinated population among those 18 and over in the respective states. Based on the studies (Randolph & Barreiro, 2020, Bartsch et al., 2020, Goldblatt et al., 2022), we use 70 percent vaccination rate as the threshold to achieve the herd immunity. We think 70 percent is a more flexible threshold which is good enough to curve the spread at least.

3.3. Predictor variables/features

We include several variables in our models that may influence the state-specific vaccination rate. We categorize the variables into four groups: economic indicators, COVID-19-related indicators, Google mobility data, and COVID-19-related policy measures. Economic indicators and the Google Mobility Report are published on a monthly and daily basis, respectively, and we aggregate them at the state level for analysis.

3.3.1. Economic features

Among the economic features, we include coincident indexes (Federal reserve bank of Philadelphia, 2020) and retail sales. The federal reserve bank of Philadelphia created the coincidence index[10] which includes four state-level specific variables; "nonfarm payroll employment, unemployment rate, average hours worked in the manufacturing sector, and wage and salary disbursements". The last variable, wage, and salary disbursement, are "deflated by the consumer price index (U.S. city average)". This index aims to use a single measure to summarize the current economic condition of the states across the U.S. Please see Stock and Watson (1989), and Crone and Clayton-Matthews (2005) for more detail on this feature.

Retail Sales data is extracted from US Census Bureau.[11] The U.S. census bureau uses three data sources, monthly retail trade survey data (MRTS), administrative data, and third-party (point-of-sales) data, to develop the monthly state retail sales (MSRS) database. MSRS includes the trend ratio (year-to-year percentage change) in monthly retail sales across the states in the United States. Notably, MSRS excludes non-store retail sales but includes eleven retail sub-sectors of NAICS, which stands for the North American Industry Classification System. The eleven sub-sectors[12] are furniture and home furnishings stores, electronics and appliance stores, building material and garden equipment and supplies dealers, food and beverage stores, health and personal care stores, clothing and clothing accessories stores, sporting goods, hobby, book, and music stores, general merchandise stores, miscellaneous store retailers and gasoline stations. As we can see, MSRS provides a comprehensive picture of the state of the U.S. economy.

3.3.2. Google mobility data

We used several variables from Google Covid-19 Community Mobility Report (CCMR). These variables measure changes for each day from the baseline, which is calculated as "the median value of the 5-week period of January 3 - February 6, 2020" (Google LLC, 2021). The

[7] JHU CSSE COVID-19 Data: https://github.com/CSSEGISandData/COVID-19

[8] Kaiser Family Foundation data: https://www.kff.org/report-section/state-covid-19-data-and-policy-actions-policy-actions/.

[9] Google COVID-19 Community Mobility Reports: https://www.google.com/covid19/mobility/.

[10] Data access:https://www.philadelphiafed.org/surveys-and-data/regional-economic-analysis/state-coincident-indexes.

[11] Data access: https://www.census.gov/retail/state_retail_sales.html.

[12] The names are taken from here: https://www.bls.gov/iag/tgs/iag_index_naics.htm based on https://www.youtube.com/watch?v=zBe6yQEJ1vQ&t=344s.

Table 2
Summary statistics of the features (Vaccination threshold met).

Features	Type	Mean	Median	Standard dev.
Economic Index	Float	127.64	124.15	11.04
Grocery Travels	Float	6.42	6.20	5.15
Covid Growth Rate	Float	0.04	0.05	0.07
Mask Mandate State	Categorical (0/1)	0.60	0.0	0.84
Mask Mandate School	Categorical (0/1)	0.80	0.0	0.42
Park Visits	Float	75.64	69.40	30.44
Political Affiliation (PA)	Categorical (0/1)	0.10	0.0	0.31
Residential Travels	Float	5.81	5.83	1.82
Retail Recreation	Float	2.51	2.22	6.90
Retail Sales	Float	37.43	33.45	7.20
State Emergency	Categorical (0/1)	0.40	0.0	0.51
Transit	Float	−15.22	−17.56	15.02
Vaccine Mandate State	Categorical (0/1)	0.90	1.0	0.31
Vaccine Mandate School	Categorical (0/1)	0.50	0.50	0.52
Workplace Travels	Float	−24.56	−25.17	3.12

Table 3
Summary statistics of the features (Vaccination threshold not met).

Features	Type	Mean	Median	Standard dev.
Economic Index	Float	131.23	130.87	13.80
Grocery Travels	Float	10.83	10.71	6.69
Covid Growth Rate	Float	0.27	0.22	0.32
Mask Mandate State	Categorical (0/1)	0.25	0.0	0.63
Mask Mandate School	Categorical (0/1)	0.47	0.0	0.71
Park Visits	Float	60.56	63.82	45.85
Political Affiliation (PA)	Categorical (0/1)	0.62	1.0	0.49
Residential Travels	Float	3.39	3.43	1.34
Retail Recreation	Float	6.90	8.75	7.20
Retail Sales	Float	31.77	31.40	3.94
State Emergency	Categorical (0/1)	0.50	0.50	0.50
Transit	Float	5.90	8.21	14.18
Vaccine Mandate State	Categorical (0/1)	0.82	1.0	0.84
Vaccine Mandate School	Categorical (0/1)	0.07	0.0	0.26
Workplace Travels	Float	−18.55	−18.42	3.05

variables we used are mobility/travels in workplaces, residential places, retail and recreation centers, parks, grocery stores & pharmacies, and transit stations such as subway, train, and bus stations.

3.3.3. *COVID-19 related features*

COVID-19-related state features include three variables. These are as follows: a total number of doses administered to people 18 years and above based on the jurisdiction where the recipient lives, the total number of doses administered to people 65 years and above based on the jurisdiction where the recipient lives, and growth rate of the daily COVID-19 cases.

3.3.4. *COVID-19 related state policy features*

COVID-19-related state policy features include a statewide emergency declaration, statewide face mask requirement, any vaccine mandate, face mask requirement in schools, and vaccine mandate for school employees. It is important to note that all the state policy measures are not the same, but they are very similar in terms of their nature and how they are enforced.

Lastly, we define the variable 'political affiliation' as the party affiliation of governors from which they got elected. The data is taken from Ballotpedia (Ballotpedia, 2020).We provide a more concise description and interpretation of the feature variables in Table 1.

4. Summary statistics

In Tables 3 and 5, we present statistical summaries of the features by vaccination threshold met or not met, respectively. We proceeded to present a visual representation of the summary statistics as well in Fig. 1

Tables 2 and 3 compare features between the group of states that met the vaccination threshold and the group of states that did not. In terms of political affiliation (1 means Republican, and 0 means Democrat), the states that met the vaccination threshold are predominantly Democrat (90%); in contrast, the majority of the states that did not meet the threshold are Republican (62%). Even though median grocery travel, residential travel, and travels to retail and recreation increases across all the states compared to the baseline numbers, the increase in the group of states that met the vaccine threshold is lower compared to the increase in the group of the state that did not. Interestingly, the median number of park visits increased in the group of states that met the threshold than the other group of states. The most substantial difference was observed in travels to transit stations (such as subway, bus, and train stations). In contrast to an increase of 8.21% from the baseline number in the states that did not meet the vaccine threshold, the median number of travels to transit stations decreased by 17.56% in the states that were able to meet the threshold. Travels to workplaces decreased in both groups of states; however, it decreased more in the states that crossed the threshold (−25.17 vs. −18.42). More states that met the threshold implemented mask mandates in schools, statewide mask mandates, declaring state emergency, and state and school vaccine mandates than the other states who fell short of meeting the threshold. Interestingly, the later group of states experienced slightly lower retail sales than the former; however, the economic index shows that the economies of the former group of states (those who met the threshold) were performing better.

5. Method

CHAID, abbreviation for Chi-squared Automatic Interaction Detection, is a Decision Tree algorithm (Xu, Zhou, G Asteris, Jahed Armaghani, & Tahir, 2019) we use for this study. The main goal of Decision ML algorithms is to split the dataset into mutually independent buckets in relation to the target variable. CHAID method uses Chi-squared measurement metrics to determine which features are the most

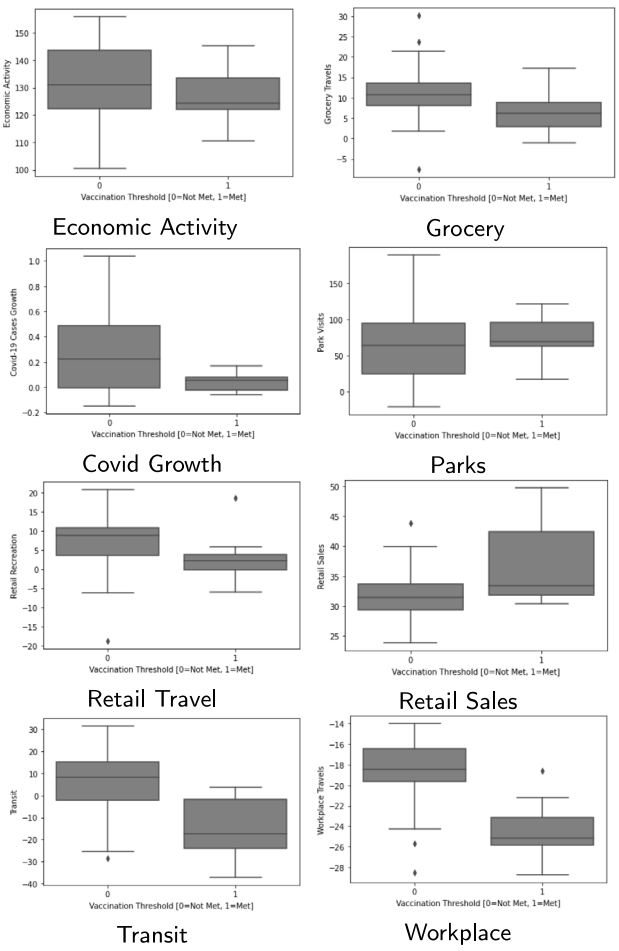


Fig. 1. Summary Statistics of Continuous Value Features by Vaccination Threshold.

important, and the procedure is applied repeatedly and recursively until the data is partitioned into mutually distinct, exhaustive subsets that best reflect the target variable (Kass, 1980).

In more technical terms, CHAID uses the following χ^2 formula to all features, and the feature with the highest value is chosen as the most significant feature, and the data is divided based on this feature first[13]:

$$\chi^2 = \sqrt{[(T - \hat{T})^2/\hat{T}]} \tag{1}$$

In this context, T represents the actual value of the target variable, which is the vaccination threshold variable. The target's anticipated value is $\hat{T}$. For example, suppose our objective is a binary variable with two classes: not meeting the threshold (0s) and meeting the threshold (1s), and we wish to use the χ^2 algorithm on the political affiliation feature. There are two classes of political affiliation: Democratic affiliation and Republican affiliation. First, we count the number of instances in our target variable where the threshold was not met (the 0s) and when the threshold was met (the 1s) for each class of the political affiliation feature. Then, for each class of the political affiliation feature, we add the 0s and 1s to get the total number of instances for each of those classes. Following that, we divide the total, in this example by two, to get the projected $\hat{T}$. We divide by two since our actual target variable T is split into two classes (0 and 1). Then all that remains is to use the above formula to calculate the χ^2 value. For each class of the

[13] A non-technical step-by-step explanation: https://sefiks.com/2020/03/18/a-step-by-step-chaid-decision-tree-example/.

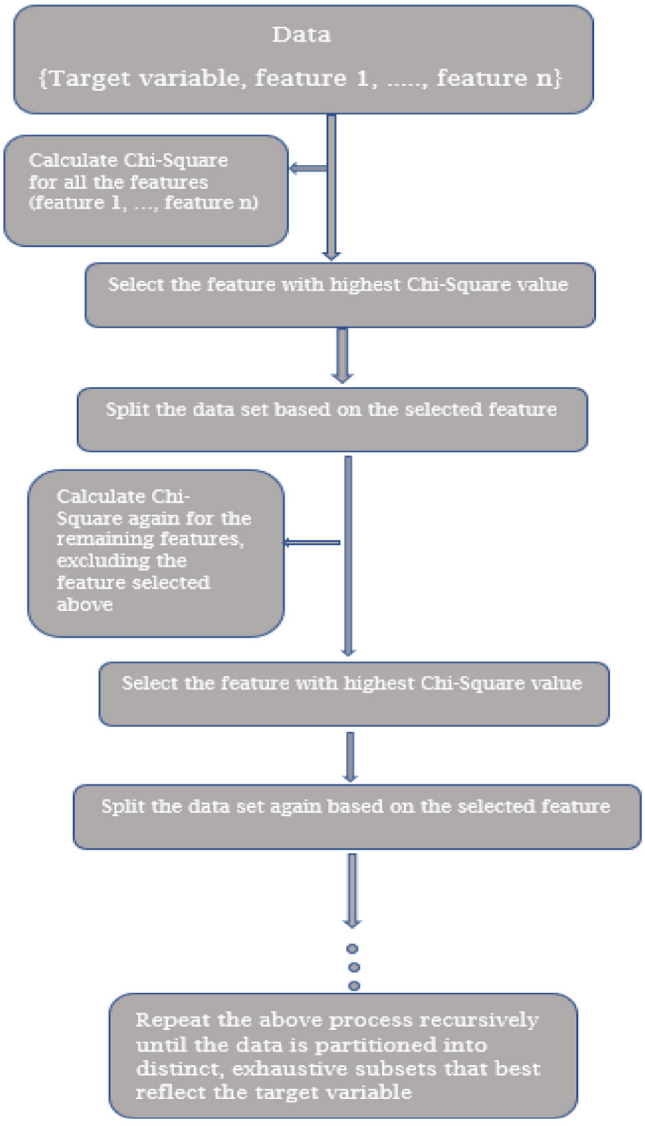


Fig. 2. CHAID model structure.

target (0 and 1), we construct different χ^2 values for each class of the political affiliation feature (democratic and republic). As a result, we end up with $\chi^2_{0,dem}$, $\chi^2_{0,rep}$, $\chi^2_{1,dem}$, and $\chi^2_{1,rep}$ in this example. After that, we add all of these χ^2 values to get a final χ^2 value for the political affiliation feature. The same procedure is repeated to the remaining features in our data, and the feature with the greatest χ^2 value is the most dominating feature, upon which the entire data is split first. The same formula is then applied to each subset of the data to identify the most significant feature again. This procedure is repeated recursively until the data is partitioned into distinct, exhaustive subsets that best reflect the target variable. The intuition behind the maximum χ^2 value feature is the following: When we examine that formula's numerator, we can see that it embeds the anticipated target variable's divergence from the actual target. Large divergences, according to intuition, are surprising and unexpected. It is possible that it is not due solely to chance, and hence features showing large χ^2 values have an important effect on predicting the target. Small χ^2 readings (small departures) are quite normal and can be attributed to chance (Narasimhan, 2013) (see Fig. 2).

CHAID enables the production of prediction models and eliminates unnecessary predictor variables (identification of the most important predictors) (Hsu & Kang, 2007). It also enables the creation of a simple

graphic (tree dendrogram) that finds mutually distinct data segments with shared attributes (Hsu & Kang, 2007). Yet another advantage of using CHAID is that it automatically removes the predictors that are not critical to the outcome. In this way, it prunes the decision tree to prevent it from overfitting (Xu et al., 2019).

We also wanted to see if the CHAID model's generated dominating factors are resilient, meaning that alternative tree-based machine learning approaches give us with the same collection of relevant variables, since the main focus of this paper is figuring out the major drivers for states attaining the vaccination threshold. We are more concerned with whether we acquire the same set of variables than with the order in which these variables are important.

Our choice of using decision tree-based model is for a combination of the following reasons: We wanted to choose an algorithm that met the technical requirements while still being readable enough to persuade non-technical stakeholders. Our research aims to quantify scores to features that represent their relative value in generating predictions. These comparative scores can indicate which features are more pertinent to the target (vaccination threshold in our case) and give policymakers and stakeholders a ranking that will benefit them in policy-making. We also wanted to use algorithms not far off the gold standard in practice, such as Neural Networks methods. Tree-based and neural network approaches are similar in that they breakdown problems gradually and try not to identify a single complex decision border capable of partitioning the whole data, like some other algorithms such as Support Vector Machines.

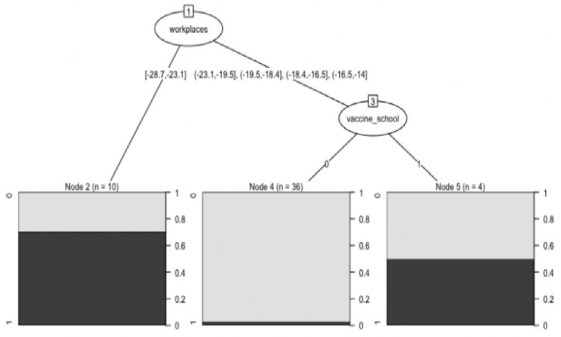
Our work, along with scholars and experts in the field, is also directed at policymakers and other stakeholders who do not often come from a technical or scientific background. Decision trees are simple to use and explain, and they show all the options in a way that makes it easy to compare with only a few short explanations. They are intuitive and think in the same way that humans do when they make a decision. A decision tree has the substantial advantage of pushing the evaluation of all conceivable outcomes of a decision by tracing each path to a conclusion. It generates a complete analysis of the effects along each route.

6. Results

6.1. Model-1

The first model defines the target variable as whether the vaccination coverage of the states' population of 18 and over meets the 70% threshold required to at least curtail the pandemic. The predictor variables are total number of doses administered to people 18 years and above based on the jurisdiction where recipient lives, total number of doses administered to people 65 years and above based on the jurisdiction where recipient lives, growth rate of the daily COVID-19 cases, coincidence index, political affiliation of the governor, whether a state declares emergency or not, statewide mask mandate, statewide vaccine mandate, mask mandate for schools, vaccine mandates for schools, retail sales and the mobility variables which includes mobility/travels in commercial and leisure, grocery and pharmacy, parks, transportation terminals, offices, and residential areas.

As Fig. 3 shows, the CHAID algorithm only kept the two most dominant predictors in the Chi-squared test out of all the predictors that passed. In the CHAID dendrogram figure above, the algorithm has determined that dividing our state-level data into 3 terminal nodes or buckets is the most predictive way. Each node corresponds to a different set of predictors. Node-2 and Node-5 predict states that will meet the vaccination threshold, whereas Node-4 predicts states that will not. With practically all gray, Node 4 appears to be the most homogeneous bucket — almost all of the states in this bucket failed to reach the vaccination threshold. There are 36 states in this group. We also get the 'error rate' from the executed algorithm, which is 2.8 percent for this node. This indicates that we were only 2.8 percent off




```
[1] root
|  [2] workplaces in [-28.7,-23.1]: 1 (n = 10, err = 30.0%)
|  [3] workplaces in (-23.1,-19.5], (-19.5,-18.4], (-18.4,-16.5], (-16.5,-14]
|  |  [4] vaccine_school in 0: 0 (n = 36, err = 2.8%)
|  |  [5] vaccine_school in 1: 0 (n = 4, err = 50.0%)
Number of inner nodes:  2
Number of terminal nodes: 3
```

Fig. 3. CHAID dendrogram for Model 1.

in our projection that these 36 states would not meet the immunization threshold. We have high confidence that this is a group of states about which we are pretty concerned that they will be unable to combat the epidemic properly, assuming that they are not achieving herd immunity in other ways. The most important takeaway from the result above is this group's shared traits, which are very relevant for policy purpose. When travels/visits to 'workplaces' fall between 23.1 percent and 14 percent below the baseline and school personnel are not required to vaccinate, a state will fall short of meeting the threshold vaccination level. Most states have met the vaccine requirement in another terminal node, Node 2. The error rate for this node is 30%, which means we are 70% correct in our prediction that the states in this group would reach the vaccination threshold required to curtail the epidemic. The criteria for this group of states are that travel/visits to 'workplaces' decreased by 28.7% to 23.1%, inclusive.

The key and possibly more qualitative message from this CHAID dendrogram is that travel/visits to workplaces and vaccine mandates for school personnel are the best predictors of passing the immunization threshold. The Chi-squared (χ^2) values for workplace travels and school vaccination mandate are 21.25 (p-value <0.001) and 7.82(p-value <0.005). We can state that trips to workplaces were the most important predictor and that vaccine mandates for school staff were the second most important predictor.

6.2. Model-2

In this specification, we remove the following COVID-19-related policy variables: Statewide emergency declaration, statewide face mask requirement, any vaccine mandate, face mask requirement in schools, and vaccine mandate for school employees and retain the remaining variables from model-1. We wanted to determine the effect of removing the actions mandated by these policies on our model (model-2). This enables us to identify more organic dominant drivers of the likelihood of states meeting the threshold.

In the CHAID dendrogram (Fig. 4), node-2 predicts states that will meet the vaccination threshold, whereas Node-4 and Node-5 predict states that will not. Like model-1, workplace travels (χ^2 = 21.25, p-value < 0.001) remain the most significant predictor of states meeting the vaccine threshold. Interestingly, another significant predictor is the governors' political affiliation (χ^2 = 6.85, p value = 0.008). We have workplace visits (WV) that divide the states into two contrasting

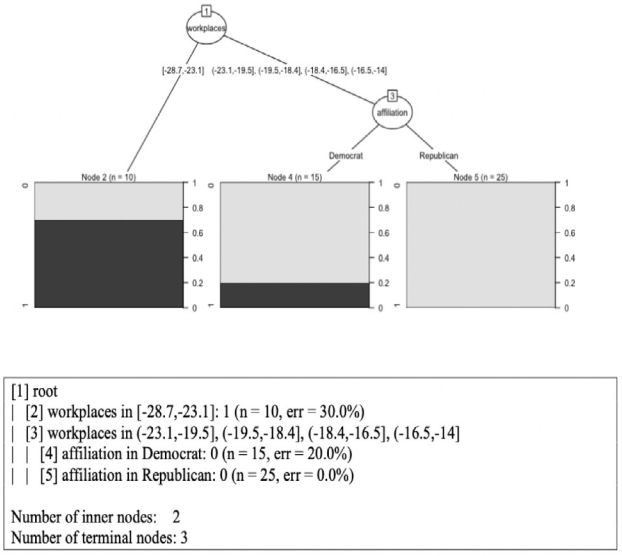


Fig. 4. CHAID dendrogram for Model 2.

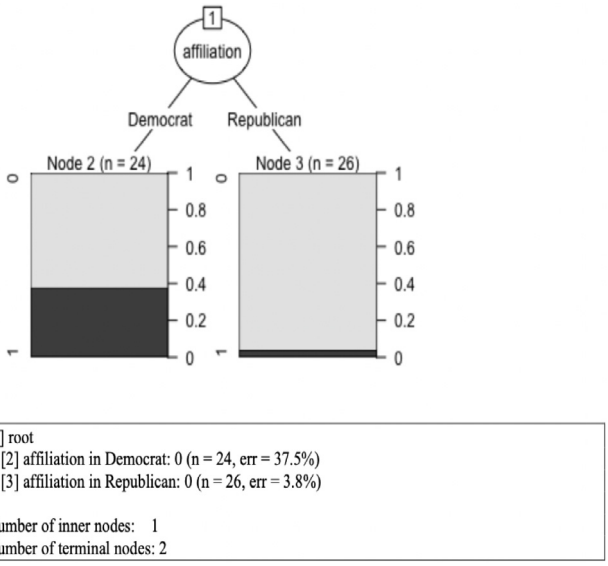


Fig. 5. CHAID dendrogram for Model 3.

segments: decreases in travel relative to the baseline of between [28.7% and 23.1%] with 70% likely to meet vaccination thresholds, and decreases in travel relative to the baseline of between (23.1% and 14%]. The latter segment is further subdivided by political affiliation (PA): states with a Democratic party affiliation are 20% likely to meet the threshold. In contrast, states with a republican party affiliation were 0% likely, which means they will never meet the threshold.

This specification provides policymakers with more actionable group characteristics, as we can derive an entirely homogeneous group of states (Node-5). We have identified a group of states that are unlikely ever to meet the threshold; they are defined by a decrease in workplace travel of between 23.1 percent and 14 percent and a Republican political affiliation.

6.3. Model-3

In model-2, we find political affiliation as a significant predictor of meeting the vaccination threshold. It was an extremely intriguing finding, and, In model-3, we investigate it further. We want to see if political affiliation remains the dominant predictor after removing mobility and/or travel-related variables from our model in this specification (model-3).

In the CHAID dendrogram (Fig. 5), both Node-2 and Node-5 predict states that will not meet the vaccination threshold. We find political affiliation to be the single most dominant predictor ($\chi^2 = 6.25$, p value = 0.008). Additionally, the prediction pattern remains consistent: states with a Republican political affiliation are only 3.8% likely (0% previously) to meet the vaccination threshold, making it more likely that the pandemic will fail to be contained. In contrast, states with a Democratic political affiliation are more likely to contain the pandemic, 37.5%.

7. Model performance

We have developed models that can segment 50 states and provide a clear view of the major drivers affecting the likelihood of attaining the vaccination threshold necessary to contain the pandemic. Now, it seems plausible to understand how good our models' predictions are. The target variable (whether it meets the 70% threshold or not) is hidden from the already trained models (model 1, model 2, and model 3) and only the features/predictor variables are fed to the models. Then, depending on the attributes passed, we let our models forecast the target/outcome variable. Finally, we compare the forecasted outcomes to the actual target outcomes that we did not disclose to our models and generate a measure of model accuracy.

Table 4
Models' performances.

Models	In-sample		Cross-validated	
	Accuracy	Sensitivity	Accuracy	Sensitivity
Model 1	0.88	0.925	0.88	0.92
Model 2	0.80	1.0	0.80	1.0
Model 3	0.80	1.0	0.80	1.0

We first evaluate the decision tree algorithm's accuracy using the same data that was used to fit the algorithm (i.e., in-sample accuracy). Then we present the k-fold cross validation (CV) results, which evaluate our model's generalized performance. We chose 10-fold CV as studies on real data show that it is a good benchmark for balancing the bias–variance tradeoff (Ron, 1995) and addresses the overfitting issue.

Our model's accuracy informs us what proportion of situations it accurately predicts. Please note that we are attempting to predict states that would not meet the threshold (in our case, the 'positive class' is 0). Here positive class does not imply that failing to achieve the vaccination threshold is a good thing; instead, it relates to what we wish to forecast, in this case failing to meet the vaccination threshold. In our instance, however, 'sensitivity' is a more essential parameter than 'accuracy'. We would like to have a small number of cases when our model predicts that a state will meet the threshold, but it will not. In other words, we want as few 'False Negatives' as possible. The sensitivity measure, defined as follows, can capture this phenomenon: Correctly predicted 0 instances / (Correctly predicted 0 cases + Incorrectly predicted 1 cases).

The Table 4 summarizes the in-sample and cross validated accuracy and sensitivity scores for various model configurations. We observe that our models perform identically in and out of sample. This offers us assurance that our algorithm is not excessively overfitting the training data and that the decision tree's prediction performance on unseen data is adequate. Again, we are more interested in the out-of-sample sensitivity scores, and all models perform admirably in this regard.

7.1. Model-1 performance

Model-1 gives an accuracy rate of 88 percent. It means Model-1 correctly predicted 44 out of 50 cases (88% of 50 states) in terms

of whether or not they would reach the vaccination threshold. The sensitivity score is 92.5%. This implies that 37 out of 40 states that did not meet the vaccination threshold were correctly predicted by the algorithm. The higher this value is, the fewer false negatives are likely to occur. In this example, our model works admirably, with a sensitivity of 92.5 percent.

7.2. Model-2 performance

Model-2 performs better than model-1, with no false negatives. As shown in Table 7, the model could predict all the states that would not meet the vaccine threshold perfectly (100%). Although the accuracy is 80%, as previously explained, this is a less desirable and/or useful metric on which to evaluate our model.

7.3. Model-3 performance

Please note that we tested Model-3 not as a candidate model but to test the predictor of political affiliation's robustness. We still report the performance of Model 3, which is identical to Model 2 with an accuracy of 80% and a 'sensitivity' metric of 100%.

This section shows that three different CHAID-based models perform well. However, the ranking of the features sometimes changes in model-1, while ranking for model-2 and model-3 was consistent over multiple runs of the CHAID algorithm. This is another major reason we presented an extensive robustness check in the next two sections. Our main goal is to explore if the features and their ranking remain consistent across multiple machine learning algorithms.

8. Robustness check-1

In this section, we use another popular tree-based machine learning algorithm, Random Forest (RF), to test the validity of our findings in model-1 and model-2. We extract the set of important variables and their ranking using the RF algorithm. A critical characteristic of random forests is that they produce measures of variable importance that may be used to find the most important predictor variables (Hapfelmeier, Hothorn, Ulm, & Strobl, 2014, Breiman, 2001). It also works well when we have a small sample size, and highly correlated sample features (Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008). Random Forest ranks the variables in terms of a 'mean decrease in accuracy' (MDA). The MDA score indicates the accuracy lost when each variable is removed from the model. The more precision is lost, the more critical the variable is for classification success. The variables are listed in order of decreasing relevance.

As we mentioned earlier, we wanted to see if the CHAID model's generated dominating factors are resilient, meaning that alternative tree-based machine learning approaches give us the same collection of relevant variables since the main focus of this paper is figuring out the major drivers for states to attain the vaccination threshold. We are more concerned with whether we acquire the same set of variables than with the order in which these variables are important.

In case of model-1, we observe that RF provides the same dominant variables (workplaces (MDA score = 0.28) and vaccine school (MDA score =0.07)) as CHAID. It also maintains the same raking. In model-2, we continue to get workplace travels and political affiliation as the primary predictors, with MDA scores of 0.19 and 0.66, respectively.

9. Robustness check-2

In this section, we provide further evidences in support of our findings in model-1, model-2 and model-3. Specifically, we test the validity of our findings that workplace travel is the most important variable in achieving the vaccination threshold. Also, we wanted to test if, after removing the COVID-19-related policies and workplace travel variables, political affiliation remains the most important feature in the model-3 specification.

To investigate this, we use non-tree-based ML algorithms and rank the features by their absolute relevance score. We apply Lasso (LCV), Ridge (RCV), ElasticNet (ECV), Logistic Regression (LR), Support Vector Machine (SVM), and XGBoost (XGB). As our initial results come from a decision tree-based model, we wanted to test the validity of our results with different class/type of ML algorithms. That said, we still apply XGB and SVM for their widespread popularity. Our results do not change even if we drop these algorithms. All the algorithms mentioned above provide different approaches to answer the question we are asking, that is, finding the most important features in predicting which states would achieve the vaccination threshold.

Least absolute shrinkage and selection operator, commonly known as LASSO (LCV), is effective in variable selection and regularization. An extension of ordinary least square (OLS), LASSO is efficient in increasing prediction accuracy by adding a penalty to the sum of residual square (RSS) (Tibshirani, 1996). Mathematically, it takes the following form:

$$min(RSS + \lambda \sum_{i=1}^{p} |\beta_i|) \tag{2}$$

Please note that the sum of the beta coefficients in the above equation does not include the intercept. The penalty depends on the value of λ: a value less than one slows down the penalty, whereas a value greater than one does the opposite.

Compared to LASSO, RIDGE regression (RCV) uses an L2 type of penalty, which keeps all the coefficients and shrunk them by the same factor. This penalty is defined as the square of the regression coefficients, which essentially shrinks the coefficients of the input variables, which are less important in predicting the output (the vaccine threshold in our case) (McDonald, 2009). Mathematically, we minimize the following equation:

$$min(RSS + \lambda \sum_{i=1}^{p} \beta_i^2) \tag{3}$$

ElasticNet (ECV) is an approach where we combine both L1 and L2 types of penalty, that is, we linearly add penalties from both LASSO (L1) and RIDGE (L2) regression to the RSS. Mathematically, we minimize the following equation (Hastie, Tibshirani, Friedman, & Friedman, 2009):

$$min(RSS + \lambda(\frac{1-\alpha}{2} \sum_{i=1}^{p} |\beta_i| + \alpha \sum_{i=1}^{p} \beta_i^2)) \tag{4}$$

In the equation above, if $\alpha = 0$ it becomes RIDGE regression and when $\alpha = 1$ it becomes LASSO regression. ElasticNet allows a better balance between the two penalties, which may improve model performance (Hastie et al., 2009).

Support vector machine (SVM) is a popular classification technique that classify inputs in a high dimensional space by building a hyper-plane (Gove & Faytong, 2012). It also uses regularization in order to avoid artifacts. Support vectors are the values that are closest to the classification margin. The purpose of SVM algorithm is to maximize this margin between the support vectors and the hyperplane. Mathematically, we maximize the following equation (Raschka, 2015):

$$\frac{\boldsymbol{w}^T(\boldsymbol{x}_{pos} - \boldsymbol{x}_{neg})}{\|\boldsymbol{w}\|} = \frac{2}{\|\boldsymbol{w}\|} \tag{5}$$

subject to

$$y^{(i)}(\boldsymbol{w}_0 + \boldsymbol{w}^T \boldsymbol{x}^{(i)}) \geq 1 \qquad \forall i \tag{6}$$

Logistic regression (LR) is useful in modeling the probability of a certain class or binary outcome (in our case, whether a state is achieving the threshold or not). In such cases, the assumption is that the given data is linearly separable. The LR model is written as (Wooldridge, 2015):

$$P(y = 1|x) = G(\beta_0 + \beta_1 x_1 + \cdots . + + \beta_k x_k) \tag{7}$$

Table 5
Robustness check: Top features (Chronologically ranked) by importance across ML algorithms.

Algorithm	Model 1	Model 2	Model 3
LCV	Workplace, Economic Index, Retail Sales	Workplace, Economic Index, Retail Sales	PA
RCV	Workplace, Vaccine School, Growth Cases	Workplace, Growth Cases, Residential	PA
ECV	Workplace, Retail Sales, Economic Index	Workplace, Retail Sales, Economic Index	PA
LR	Vaccination School, PA, Workplace	PA, Growth Cases, Workplace	PA
SVM	Workplace, Vaccine Mandate, PA	Workplace, Residential, PA	PA
XGB	Workplace, Vaccine Schools, Parks	Workplace, Parks, Growth Cases	PA

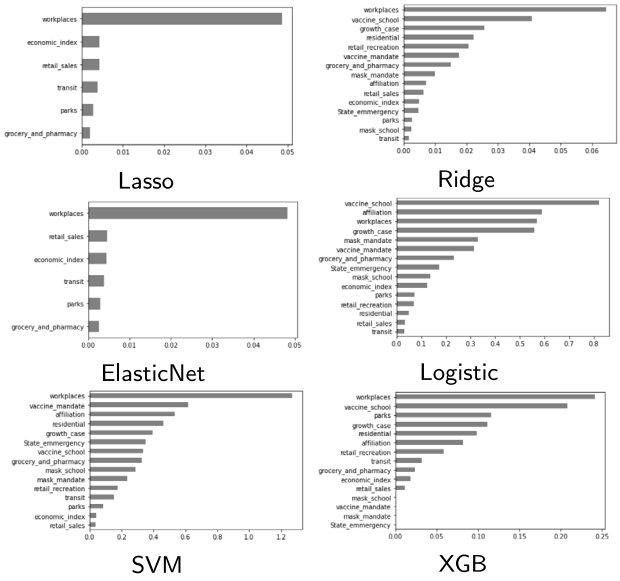


Fig. 6. Model 1 (Full Model) Feature Importance Scores (X-axis)across Algorithms.

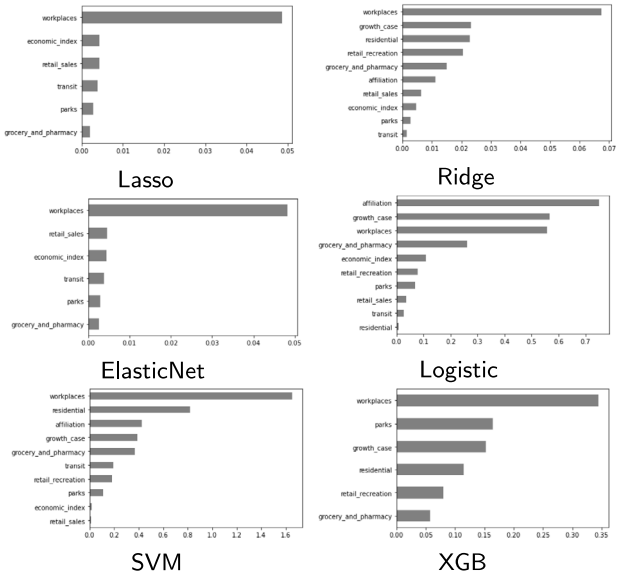


Fig. 7. Model 2 Feature Importance Scores (X-axis)across Algorithms.

where the value of the function G is strictly between 0 to 1. As a result, the estimated probability is also between 0 to 1.

Lastly, XGBoost (XGB) is a well-known classification and regression predictive modeling algorithm. It avoids overfitting problems by integrating regularization with k times iteration. The objective function that is optimized (Liang, Luo, Zhao, & Wu, 2020) is:

$$O(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{1}^{k} \lambda(f_k) + c \tag{8}$$

Here, $\sum_i^n l(y_i, \hat{y}_i)$ is the loss function, $\sum_1^k \lambda(f_k)$ is the regularization term and c is the constant. The regularization term can be written as follows:

$$\lambda(f_k) = \delta H + \frac{1}{2}\psi \sum_{1}^{T} w_j^2 \tag{9}$$

where T stands for the number of leaves, δ represents the complexity, and ψ is the penalty parameter (Liang et al., 2020).

In Table 5, we list the top three features in chronological order for model 1 and 2, and only the top feature for model 3 because the purpose of this specification was to test if political affiliation is the top predictive feature amid economic and retail sales feature. Figs. 6–8 depict the feature importance for model-1, model-2 and model 3 specifications, respectively, across the alternative ML algorithms.

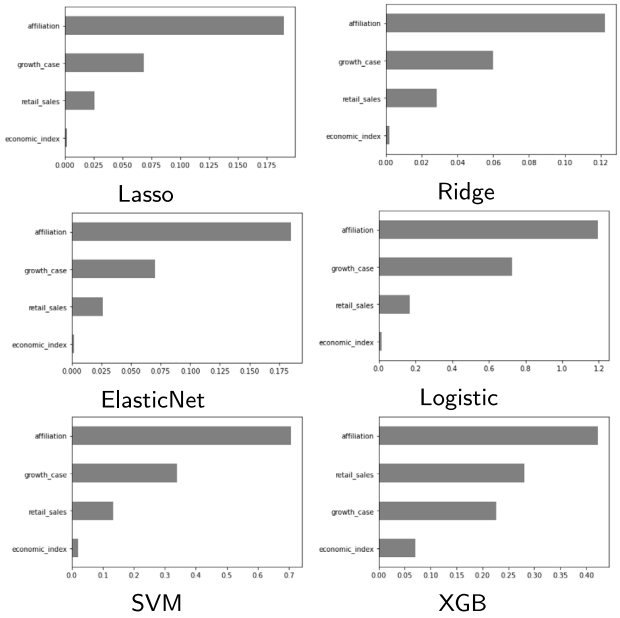


Fig. 8. Model 3 Feature Importance Scores (X-axis)across Algorithms.

9.1. *Statistical significance*

To summarize this robustness in a single number, we compute a non-parametric exact p-value similar to those found in Fisher (1935) and Abadie, Diamond, and Hainmueller (2010) (see Table 6).

By utilizing machine learning methods that are not tree-based, we can determine whether our discovery of the workplace travels feature as the top predictive feature from the CHAID decision tree algorithm is robust to algorithm heterogeneity. We order the features chronologically for each algorithm applied for model 1 and model 2 and then count the number of times workplace travel is not included in the top three features. The non-parametric p-value is then calculated as the proportion of times workplace travel does not rank among the top three in all possible methods. If this p-value is smaller than the customary 5%,

Table 6
Statistical significance of features across models.

Features	Model 1	Model 2	Model 3
Growth Case	0.83 (6)	0.5(2)	1.0 (2)
Economic Index	0.67 (3)	0.67 (3)	1.0 (2)
Political Affiliation	0.67 (3)	0.67 (3)	**0.0 (1)***
Retail Recreation	1.0 (9)	1.0 (8)	–
Grocery and Pharmacy	1.0 (9)	1.0 (8)	–
Parks	0.83 (6)	0.83 (7)	–
Transit	1.0 (9)	1.0 (8)	–
Workplaces	**0.0 (1)***	**0.0 (1)***	–
Residential	1.0 (9)	0.67 (3)	-
State Emergency	1.0 (9)	–	–
Mask Mandate	1.0 (9)	–	–
Vaccine Mandate	0.83 (6)	–	–
Mask School	1.0 (9)	–	–
Vaccine School	0.5 **(2)**	–	–
Retail Sales	0.67 **(3)**	0.67 (3)	1.0 (2)

Note: "–" indicates that these variables were not included in the respective models, and the parenthesis associated to the p-values show the relative ranking of the features in terms of the p-value.
*Denotes significance at the 1%.

Table 7
Algorithm performance comparison.

Algorithm	MAE	RMSE	EV	R^2	MSE
LCV	0.26	0.33	0.46	0.44	0.12
RCV	0.25	0.32	0.57	0.55	0.12
ECV	0.26	0.33	0.46	0.45	0.12
LR	0.18	0.35	0.50	0.72	0.18
SVM	0.18	0.37	0.35	0.52	0.18
XGB	0.20	0.36	0.70	0.84	0.20

we reject the assertion that workplace travel is not the most essential attribute. Similarly, for model 3, we calculate a non-parametric *p*-value for the political affiliation feature. In Table 1, we provide the p-values for all the features across different model specifications.

Table 1 shows that in both our comprehensive models (model-1 and model-2), workplace travels rank first with a *p*-value of 0.0, which is significant not only at the 5% level but also at the 1% level. This supports our initial claim that workplace visits are the most important predictor of meeting the vaccination threshold. The second-highest ranked feature in model-1 is vaccination mandates in schools. This confirms our original claim about this feature. In model-1, political affiliation is now tied for third place with economic index and retail sales. So, as previously stated, we combined only these identical ranked features in model-3. We also include the growth rate of COVID-19 cases because it came in second in model 2. With a *p*-value of 0.0, we find that political affiliation is the most predictive of these features, proving our original claim: In predicting vaccine threshold, political affiliation outweighs other economic and growth case variables.

Fig. 9 is really supplementing Table 5. Fig. 9 depicts the distribution of feature p-values across different model settings. We demonstrate that workplace travel is distinct from all other features in that it has the lowest non-parametric *p*-value of 0.0 across models 1 and 2. As a result, it is the most important predictor of the vaccination threshold. It further confirms our point: the workplace remains the most important predictor across multiple ML algorithms.

10. Algorithm performance comparison

The performance of alternative algorithms are compared in this section. As we are comparing the feature importance scores of these algorithms, we wanted to determine whether their performances are comparable or not too dissimilar. We employ mean absolute error (MAE), and root mean squared error (RMSE) to analyze the performances following (Brownlee, 2021) and Friedman (2001). MAE denotes the mean of the absolute discrepancies between the actual and predicted

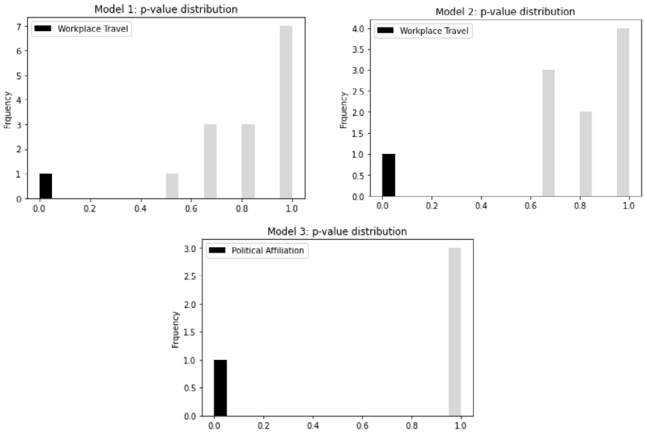


Fig. 9. Distribution of p-values for features across model 1, model 2 and model 3.

targets. RMSE is the mean of the root-squared discrepancies between the true and predicted targets. These metrics can be represented by following formulae:

$$EV = 1 - \frac{Var(T_i - \hat{T}_i)}{Var(T_i)} \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(T_i - \hat{T}_i)^2}{\sum_{i=1}^{N}(T_i - \bar{T}_i)^2} \tag{11}$$

$$MSE = (1/N)\sum_{i=1}^{N}(T_i - \hat{T}_i)^2 \tag{12}$$

$$MAE = (1/N)\sum_{i=1}^{N}|T_i - \hat{T}_i| \tag{13}$$

$$RMSE = \sqrt{(1/N)\sum_{i=1}^{N}(T_i - \hat{T}_i)^2} \tag{14}$$

Here, T_i and $\hat{T}_i$ are the true value and the predicted value of the target variable respectively for example i in the data. N is the total number of examples/observations in the data. A large portion of our alternative class of algorithms is based on regression, which conducts prediction by forecasting a numerical value. As with our original decision tree-based algorithms, we are unable to quantify the performance of these algorithms using accuracy. Having stated that, we required a single performance metric to determine whether the performance of all alternative algorithms falls within a reasonable range. We employ the k-fold cross validation technique to determine our algorithms' generalization error (performance on unknown data) credibly. This provides an acceptable trade-off between bias and variance when evaluating our models (Raschka & Mirjalili, 2017). As mentioned earlier, our goal in this paper is to identify the most important predictors of vaccination threshold at the state level so that policymakers and/or stakeholders can use them as policy tools. Deep Neural Networks and other Deep Learning algorithms are extremely complicated, and the parameterization in the hidden layers is nearly impossible to decipher. Even specialists are unable to explain the data-generating process modeled by these algorithms and extract single scores for feature relevance, which is why they are referred to as "black-box" algorithms (Lantz, 2019). Our work aims to equip policymakers and stakeholders with tools to help them design policies. We need better explainable ML techniques like Decision Tree, rather than black-box methods, so stakeholders have more visibility into the algorithm.

We present the MAE, RMSE, EV, R^2 and MSE for all alternative approaches in Table 7. We may note that the performances of the most widely used and well-established performance metrics for regression algorithms (Brownlee, 2021 and Friedman, 2001) – MAE and RMSE

– are comparable. MSE, simply the RMSE squared, is an additional often employed metric. MAEs range from 0.18 to 0.26, whereas RMSEs range from 0.33 to 0.36.EVs and R^2s range from 0.35 to 0.46 and 0.44 to 0.52, respectively, with the minor exception of XGB, where EVs are 0.70 and R^2s are 0.84. Again, we observe a small range between 0.12 and 0.20 for MSE. The non-parametric *p*-value we determined earlier embeds algorithm heterogeneity when ranking the features. We would like to see our top predictor be robust across algorithms and algorithms that perform similarly on unseen data. We evaluate the data modeling ability and/or generalization performance of these algorithms in Table 7 by employing performance measures. This really supports and strengthens our p-value-based claim regarding the most important feature. Not only are our top predictors of workplace travel (model-1 and model-2) and political affiliation (model-3) robust across algorithms, but they are also resilient across algorithms with identical generalization errors. In other words, a feature that ranks first across all algorithms with similar prediction performance lends that feature additional credibility.

11. Discussion

Taking a vaccine is primarily a personal decision. There is rich literature investigating factors affecting individual vaccination decisions. Most of these studies focus on socioeconomic and demographic factors, individuals' religious and political beliefs, attitudes towards science, source of information, and media framing. As an individual's attitude and beliefs cannot be changed overnight, it is important to understand the state and federal level policy measures that could play an important role in increasing the vaccination rate. To examine this, we use three different model specifications and several ML algorithms. The first model (model-1) includes all the state-level aggregate features mentioned in Section 3, model-2 keeps all the features of model-1 except COVID-19-related policy variables (statewide emergency declaration, statewide face mask requirement, any vaccine mandate, face mask requirement in schools, and vaccine mandate for school employees), and lastly, model-3 includes features of model-2 excluding mobility and/or travel related variables. The uniqueness of model-1 is that it helps us to understand what state-level features are most important in achieving the vaccination threshold in the presence of all the features. Compared to model-1, model-2 reveals what features are important in the absence of any COVID-19-related policy measures. Lastly, model-3 shows which feature is the most important among the economic features, the growth rate of COVID-19 cases, and the governor's political affiliation. Each of these three models offers different perspectives for the policymakers in terms of understanding the important features given the particular context of a state. All three models perform very well: in-sample accuracy ranges from 80%–88% and sensitivity is between 92.5%–100%. Please note that, in our study, 'sensitivity' is a more essential parameter than 'accuracy' because we are predicting states that would not meet the vaccination threshold ('positive class' is 0). The Section 7 section includes a detailed discussion on both these models' in-sample and out-of-sample performances.

In this study, we find that workplace travel has the most predictive power in classifying which states would meet the vaccination threshold required to achieve herd immunity among all the state-level features and policies. The following two critical variables are the vaccine mandate for school employees and the governor's political affiliation of each state. Political affiliation (PA) becomes important for many reasons. Governors hold the executive power in their respective states, which plays a critical role in COVID-19-related policies. They have the authority to formulate and enforce lockdown policies, state emergencies, vaccine mandates, face mask mandates, and many other policies. Studies (Guy Jr et al., 2021; Guzzetta et al., 2020; Joo et al., 2021; Karaivanov, Lu, Shigeoka, Chen, & Pamplona, 2021; Reiss & Caplan, 2020) show that all these policies are instrumental to curve the spread of the virus, reduce COVID-19 related deaths and hospitalization. Our results show that the political affiliation of the governors also plays a critical role in the vaccination rate in their respective states. Apart from the major policies mentioned above, the governor's role is also important in resource mobilization. As far as we know, this is a critical aspect absent from the existing studies. For example, the Alabama governor recently announced that she is allocating $400 million dollars from the coronavirus relief resources for the construction of new prisons (Duster and Valencia, 2021) even though Alabama ranked among the slowest states in administering doses per 100,000 people and the overall vaccination rate. On the other hand, even though New York is among the states with higher vaccination rates (75% of the total population as of today), additional efforts by the governor's office include a $15 million grant (New York State, 2021) to encourage immunization in neighborhoods excessively impacted by the COVID-19 outbreak.

One of the critical factors from our results is the vaccine mandate in schools. According to the Kaiser Family Foundation data, eleven states introduced school vaccine mandates. All these states are run by Democratic governors with vaccination rates above the national average. The policy of school vaccine mandate could also be a proxy variable to measure the overall seriousness of the states in combating the pandemic. However, this claim is subject to further empirical investigation.

Lastly, our results show that the variable with the most predictive power is workplace travels. A drop in workplace travel can happen for primarily two reasons: people are now traveling to and from workplaces lower than the baseline numbers, and the impact of social distance measures taken by the states and local governments. This drop in workplace travel could be viewed as a proxy variable of the effectiveness of the state and local administration's policies and increased awareness among the employers and employees who make necessary arrangements for remote work and decentralized office-oriented work style to follow public health measures. Obviously, one can expect that these population groups are more sensitive to COVID-19, which, coupled with their policy environment, plays an instrumental role in getting vaccinated at an increasing rate.

12. Conclusion

Achieving herd immunity through mass vaccination is one of the policy priorities to combat the COVID-19 pandemic. We use three different model specifications and applied Chi-squared Automatic Interaction Detection (CHAID), a Decision Tree machine learning algorithm, to identify the state-level features that predict which states would achieve the vaccination threshold necessary to develop herd immunity. We discovered that workplace travel is the most important feature to achieving the vaccination threshold and that political affiliation is the most important feature when compared to economic predictors and the growth rate of COVID-19 cases in a more conservative feature set. This is significant since economic, business (sales), and infection growth rates are the most often discussed topics in the media and public arena. The fact that political affiliation outperforms them demonstrates how politics may be a powerful predictor in these scenarios. As a robustness check, we also use several machine learning algorithms such as Random Forest (RF), Lasso (LCV), Ridge (RCV), XGBoost (XGB), ElasticNet (ECV), Logistic Regression (LR), and Support Vector Machine (SVM). Our results hold across all these algorithms; that is, workplace travels, political affiliation of the governor, and the vaccine mandate in schools remain the top three features.

One limitation of our study is that we could not capture the real-time changes in the job market movement related to workplace travels. Due to the pandemic, the U.S job market went through structural changes and experienced some unusual movements in job loss and job creation. On the other hand, even though our data includes COVID-19-related policy variables that are directly related to the governor's political affiliation, we were unable to empirically pin down all the

possible channels of the effect of the governor's political affiliation. For example, data on how the governor's office mobilizes resources to deal with the pandemic and how efficiently the states enforce the COVID-19-related policies is not observable. Future studies may address these issues along with undertaking a study that would be able to provide a causal interpretation of the results.

CRediT authorship contribution statement

Syed Muhammad Ishraque Osman: Conceptualization, Methodology, Visualization, Supervision, Writing – reviewing and editing. **Ahmed Sabit:** Conceptualization, Data curation, Writing – original draft, Validation, Writing – reviewing and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493–505.

Ballotpedia (2020). Status of lockdown and stay-at-home orders in response to the coronavirus (COVID-19) pandemic,2020. https://ballotpedia.org/Status_of_lockdown_and_stay-at-home_orders_in_response_to_the_coronavirus_(COVID-19)_pandemic,_2020#Orders_by_governor_party_affiliation. (Accessed 01 September 2021).

Bartsch, S. M., O'Shea, K. J., Ferguson, M. C., Bottazzi, M. E., Wedlock, P. T., Strych, U., et al. (2020). Vaccine efficacy needed for a COVID-19 coronavirus vaccine to prevent or stop an epidemic as the sole intervention. *American Journal of Preventive Medicine*, *59*(4), 493–503.

Borchering, R. K., Viboud, C., Howerton, E., Smith, C. P., Truelove, S., Runge, M. C., et al. (2021). Modeling of future COVID-19 cases, hospitalizations, and deaths, by vaccination rates and nonpharmaceutical intervention scenarios—United tates, April–September 2021. *Morbidity and Mortality Weekly Report*, *70*(19), 719.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brownlee, J. (2021). Regression metrics for machine learning. https://machinelearningmastery.com/regression-metrics-for-machine-learning/. (Accessed 26 June 2022).

Burch, A. E., Lee, E., Shackelford, P., Schmidt, P., & Bolin, P. (2022). Willingness to vaccinate against COVID-19: Predictors of vaccine uptake among adults in the US. *Journal of Prevention*, *43*(1), 83–93.

CDC (2022). COVID-19 vaccinations in the United States,jurisdiction. https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc. (Accessed 20 July 2021).

Chauhan, V. K., Dahiya, K., & Sharma, A. (2019). Problem formulations and solvers in linear SVM: A review. *Artificial Intelligence Review*, *52*(2), 803–855.

Cheong, Q., Au-Yeung, M., Quon, S., Concepcion, K., Kong, J. D., et al. (2021). Predictive modeling of vaccination uptake in US counties: A machine learning–based approach. *Journal of Medical Internet Research*, *23*(11), Article e33231.

Crone, T. M., & Clayton-Matthews, A. (2005). Consistent economic indexes for the 50 states. *The Review of Economics and Statistics*, *87*(4), 593–603.

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, *20*(5), 533–534.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, *35*(5–6), 352–359.

Dror, A. A., Eisenbach, N., Taiber, S., Morozov, N. G., Mizrachi, M., Zigron, A., et al. (2020). Vaccine hesitancy: The next challenge in the fight against COVID-19. *European Journal of Epidemiology*, *35*(8), 775–779.

Duster and Valencia (2021). Alabama governor defends plan to use COVID-19 relief funds to build prisons. https://www.cnn.com/2021/09/29/politics/alabama-governor-kay-ivey-covid-relief-prisons/index.html. (Accessed 20 September 2021).

Dyer, O. (2021). COVID-19: Unvaccinated face 11 times risk of death from delta variant, CDC data show. *British Medical Journal Publishing Group*.

Federal reserve bank of Philadelphia (2020). State coincident indexes. https://www.philadelphiafed.org/surveys-and-data/regional-economic-analysis/state-coincident-indexes. (Accessed 27 September 2021).

Fisher, R. (1935). *Design of Experiments* (sl ed.). Edinburgh: Oliver.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 1189–1232.

Goldblatt, D., Fiore-Gartland, A., Johnson, M., Hunt, A., Bengt, C., Zavadska, D., et al. (2022). Towards a population-based threshold of protection for COVID-19 vaccines. *Vaccine*, *40*(2), 306–315.

Google LLC (2021). Google COVID-19 community mobility reports. https://www.google.com/covid19/mobility/. (Accessed 01 September 2021).

Gove, R., & Faytong, J. (2012). Machine learning and event-based software testing: Classifiers for identifying infeasible GUI event sequences. In *Advances in computers, vol. 86* (pp. 109–135). Elsevier.

Guy Jr, G. P., Lee, F. C., Sunshine, G., McCord, R., Howard-Williams, M., Kompaniyets, L., et al. (2021). Association of state-issued mask mandates and allowing on-premises restaurant dining with county-level COVID-19 case and death growth rates—United States, March 1–December 31, 2020. *Morbidity and Mortality Weekly Report*, *70*(10), 350.

Guzzetta, G., Riccardo, F., Marziano, V., Poletti, P., Trentini, F., Bella, A., et al. (2020). The impact of a nation-wide lockdown on COVID-19 transmissibility in Italy. arXiv preprint arXiv:2004.12338.

Hapfelmeier, A., Hothorn, T., Ulm, K., & Strobl, C. (2014). A new variable importance measure for random forests with missing data. *Statistics and Computing*, *24*(1), 21–34.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction, vol. 2*. Springer.

Hsu, C. H., & Kang, S. K. (2007). CHAID-based segmentation: International visitors' trip characteristics and perceptions. *Journal of Travel Research*, *46*(2), 207–216.

Huggins (2021). Ohio ends incentive lottery with mixed vaccination results. https://apnews.com/article/ohio-coronavirus-vaccine-coronavirus-pandemic-oddities-health-8728927905ffaed3d462a274524f9e30. (Accessed 10 September 2021).

Joo, H., Miller, G. F., Sunshine, G., Gakh, M., Pike, J., Havers, F. P., et al. (2021). Decline in COVID-19 hospitalization growth rates associated with statewide mask mandates—10 states, March–October 2020. *Morbidity and Mortality Weekly Report*, *70*(6), 212.

Karaivanov, A., Lu, S. E., Shigeoka, H., Chen, C., & Pamplona, S. (2021). Face masks, public policies and slowing the spread of COVID-19: Evidence from Canada. *Journal of Health Economics*, *78*, Article 102475.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, *29*(2), 119–127.

Khubchandani, J., Sharma, S., Price, J. H., Wiblishauser, M. J., Sharma, M., & Webb, F. J. (2021). COVID-19 vaccination hesitancy in the United States: A rapid national assessment. *Journal of Community Health*, *46*(2), 270–277.

Klosterman, S. (2019). *Data science projects with Python: A case study approach to successful data science projects using Python, pandas, and scikit-learn*. Packt Publishing Ltd.

Lantz, B. (2019). *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd.

Liang, W., Luo, S., Zhao, G., & Wu, H. (2020). Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms. *Mathematics*, *8*(5), 765.

Lincoln, T. M., Schlier, B., Strakeljahn, F., Gaudiano, B. A., So, S. H., Kingston, J., et al. (2022). Taking a machine learning approach to optimize prediction of vaccine hesitancy in high income countries. *Scientific Reports*, *12*(1), 1–12.

Malik, A. A., McFadden, S. M., Elharake, J., & Omer, S. B. (2020). Determinants of COVID-19 vaccine acceptance in the US. *EClinicalMedicine*, *26*, Article 100495.

McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 93–100.

Mewhirter, J., Sagir, M., & Sanders, R. (2022). Towards a predictive model of COVID-19 vaccine hesitancy among American adults. *Vaccine*, *40*(12), 1783–1789.

Narasimhan, R. (2013). What is the most intuitive explanation for the chi square test?. https://www.quora.com/What-is-the-most-intuitive-explanation-for-the-chi-square-test. (Accessed 27 February 2022).

New York State (2021). Governor Cuomo announces allocation of $15 million to promote vaccination in communities disproportionately affected by COVID-19 pandemic. https://www.governor.ny.gov/news/governor-cuomo-announces-allocation-15-million-promote-vaccination-communities. (Accessed 12 September 2021).

Prajwala, T. (2015). A comparative study on decision tree and random forest using R tool. *International Journal of Advanced Research in Computer and Communication Engineering*, *4*(1), 196–199.

Randolph, H. E., & Barreiro, L. B. (2020). Herd immunity: Understanding COVID-19. *Immunity*, *52*(5), 737–741.

Ranstam, J., & Cook, J. (2018). LASSO regression. *Journal of British Surgery*, *105*(10), 1348.

Raschka, S. (2015). *Python machine learning*. Packt publishing ltd.

Raschka, S., & Mirjalili, V. (2017). *Python machine learning: Machine learning and deep learning with python* (2nd ed.). Scikit-Learn, and TensorFlow.

Reiss, D. R., & Caplan, A. L. (2020). Considerations in mandating a new COVID-19 vaccine in the USA for children and adults. *Journal of Law and the Biosciences*.

Ron, K. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the fourteenth international joint conference on artificial intelligence, 1995, vol. 2, no. 12* (pp. 1137–1143). American Association for Artificial Intelligence.

Ruiz, J. B., & Bell, R. A. (2021). Predictors of intention to vaccinate against COVID-19: Results of a nationwide survey. *Vaccine*, *39*(7), 1080–1086.

Sabit, A., Ahmad, S., & Abdul Baten, R. B. (2022). *Impact of State Incentives on COVID-19 Vaccination Uptake in the U.S. Health Management*. Policy and Innovation (www.HMPI.org), Volume 7, Issue 3.

Sagi, O., & Rokach, L. (2021). Approximating XGBoost with an interpretable decision tree. *Information Sciences*, *572*, 522–542.

Starr, S. (2012). Survey research: We can do better. *Journal of the Medical Library Association: JMLA*, *100*(1), 1.

Stock, J. H., & Watson, M. W. (1989). New indexes of coincident and leading economic indicators. *NBER Macroeconomics Annual*, *4*, 351–394.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 1–11.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*(1), 267–288.

Tomaschek, F., Hendrix, P., & Baayen, R. H. (2018). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics*, *71*, 249–267.

Viswanath, K., Bekalu, M., Dhawan, D., Pinnamaneni, R., Lang, J., & McLoud, R. (2021). Individual and social determinants of COVID-19 vaccine uptake. *BMC Public Health*, *21*(1), 1–10.

Walkey, A. J., Law, A., & Bosch, N. A. (2021). Lottery-based incentive in Ohio and COVID-19 vaccination rates. *Jama*, *326*(8), 766–767.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage learning.

Xu, H., Zhou, J., G Asteris, P., Jahed Armaghani, D., & Tahir, M. M. (2019). Supervised machine learning techniques to the prediction of tunnel boring machine penetration rate. *Applied Sciences*, *9*(18), 3715.

Yan, E., Lai, D. W., Ng, H. K., & Lee, V. W. (2022). Predictors of COVID-19 actual vaccine uptake in Hong Kong: A longitudinal population-based survey. *SSM-Population Health*, Article 101130.